Large Deviations for Resampling Methods and Simulations

FINAL PROGRESS REPORT

October 31, 1999

U. S. Army Research Office

DAAH04-96-1-0070

Old Dominion University Research Foundation
P.O.Box 6369, 800 West 46th Street
Norfolk, Virginia 23508.

Approved For Public Release;
Distribution Unlimited.

REPORT DOCUMENTATION PAGE

Form Approved
OMB NO. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	10/30/99	Final Report - 5/1/96 Through 10/30/99

4. TITLE AND SUBTITLE
Large Deviations for Resampling Methods and Simulations

5. FUNDING NUMBERS
DAAH04-96-1-0070

6. AUTHOR(S)
N. R. Chaganty, Principal Investigator

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(ES)
Old Dominion University Research Foundation
800 West 46th Street
Norfolk, VA 23508

8. PERFORMING ORGANIZATION REPORT NUMBER
241601

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
ARO 32886.11-MA

11. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12 b. DISTRIBUTION CODE
N/A

13. ABSTRACT *(Maximum 200 words)*

In this research project we studied two important problems in probability and statistics. First, we have established the large deviation principle for a sequence of probability measures defined on a product space, when the marginal and the conditional distributions possess the large deviation property. We have used the result to study the large deviation behavior of the bootstrap resampling procedure and robustness of location parameter tests in contaminated normal populations via Bahadur slopes and efficiencies. The second problem is concerned with the statistical analysis of longitudinal data. In recent years the GEE method has become a popular tool for analyzing discrete longitudinal data. The method uses a generalized quasi-score function to estimate the regression parameter, and moment estimates for the correlation parameter. Despite the popularity, the GEE method has some inherent pitfalls. In this research we have developed an alternative estimation procedure which overcomes those pitfalls. This alternative method is known as the Quasi-least squares (QLS), since it uses a partial minimization based on the principle of (generalized) least squares. We have shown that the QLS estimates are feasible, consistent and asymptotically normal.

14. SUBJECT TERMS
SEE ATTACHED

15. NUMBER IF PAGES
127

16. PRICE CODE
N/A

17. SECURITY CLASSIFICATION OR REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

Large Deviations for Resampling Methods and Simulations

FINAL PROGRESS REPORT

October 31, 1999

U. S. Army Research Office

DAAH04-96-1-0070

Old Dominion University Research Foundation
P.O.Box 6369, 800 West 46th Street
Norfolk, Virginia 23508.

Approved For Public Release;
Distribution Unlimited.

Large Deviations for Resampling Methods and Simulations

FINAL PROGRESS REPORT

October 31, 1999

U. S. Army Research Office

DAAH04-96-1-0070

Old Dominion University Research Foundation
P.O.Box 6369, 800 West 46th Street
Norfolk, Virginia 23508.

Approved For Public Release;
Distribution Unlimited.

Table of Contents

1 Statement of the problems studied.

In this research project we studied two very important problems in probability and statistics. The first one is a problem in the theory of large deviations with applications to studying robustness of statistical procedures, efficiencies and the bootstrap resampling method. More specifically, we have established the large deviation principle for a sequence of probability measures $\{\mu_n\}$ on a product space $\Omega_1 \times \Omega_2$ when the corresponding sequences of marginal and conditional distributions possess the large deviation property. We have used this result to study the large deviation behavior of the bootstrap resampling procedure. And also used the result to study robustness of location parameter tests in contaminated normal populations via Bahadur slopes and efficiencies. The second problem is concerned with the statistical analysis of longitudinal data. In a seminal paper Liang & Zeger (1986, *Biometrika* **73**, 13-23) introduced the generalized estimating equations (GEE) as a statistical tool for analyzing longitudinal data. The GEE method uses a generalized quasi-score function to estimate the regression parameter, and moment estimates for the correlation parameters. Recently, Crowder (1995, *Biometrika*, **82**, 407-410) has pointed out some pitfalls with the estimation of the correlation parameters in the GEE method. In this research we developed an alternative estimation procedure which overcomes those pitfalls. This alternative method is known as the Quasi-least squares (QLS) since it uses a partial minimization, based on the principle of (generalized) least squares. Below we will give a brief outline of the technical details of the work done under this contract.

2 Summary of the most important results.

2.1 Large deviations for joint distributions.

Let Ω be a Polish space, that is, a complete separable metric space and $\mathcal{B}$ be the Borel σ-field on Ω containing all the open and closed subsets of Ω. A function $I(x) : \Omega \to [0, \infty]$ is said to be a *rate function* if it is lower semi-continuous. Let $\{\mu_n\}$ be a sequence of

probability measures on $(\Omega, \mathcal{B})$. We say that $\{\mu_n\}$ obeys *large deviation principle* (LDP) with rate function $I(x)$ if the following conditions are satisfied:

$$(1) \qquad \limsup_n \frac{1}{n} \log \mu_n(C) \leq -I(C)$$

$$(2) \qquad \liminf_n \frac{1}{n} \log \mu_n(G) \geq -I(G)$$

for all closed sets C and for all open sets G of Ω. The rate function $I(x)$ is known as a *proper rate function* if for each $L \geq 0$, the level set $\{x : I(x) \leq L\}$ is a compact subset of Ω. Let $(\Omega_1, \mathcal{B}_1)$, $(\Omega_2, \mathcal{B}_2)$ be two Polish spaces with their associated Borel σ-fields. Let $\{\mu_{1n}\}$ be a sequence of probability measures on $(\Omega_1, \mathcal{B}_1)$ and $\{\nu_n(x_1, B_2)\}$ be a sequence of transition functions on $\Omega_1 \times \mathcal{B}_2$. Consider a sequence of probability measures $\{\mu_n\}$ on the product space $(\Omega, \mathcal{B}) = (\Omega_1 \times \Omega_2, \mathcal{B}_1 \otimes \mathcal{B}_2)$ given by

$$\mu_n(B_1 \times B_2) = \int_{B_1} \nu_n(x_1, B_2)\, d\mu_{1n}(x_1)$$

for $B_i \in \mathcal{B}_i$, $i = 1, 2$. We say that the sequence of probability transition functions $\{\nu_n(x_1, \cdot), x_1 \in \Omega_1\}$ satisfies the LDP *continuously in* x_1 with rate function $J(x_1, x_2)$, or simply LDP *continuity condition* holds, if

(i) For each $x_1 \in \Omega_1$, $J(x_1, \cdot)$ is a proper rate function on Ω_2.

(ii) For any sequence $\{x_{1n}\}$ in Ω_1 such that $x_{1n} \to x_1$, the sequence of measures $\{\nu_n(x_{1n}, \cdot)\}$ on Ω_2 obeys the LDP with rate function $J(x_1, \cdot)$, and

(iii) $J(x_1, x_2)$ is jointly lower semi-continuous in (x_1, x_2).

Our main result can be stated as follows. Suppose that the sequence $\{\mu_{1n}\}$ obeys the LDP with proper rate function $I_1(x_1)$ and the sequence of probability transition functions $\{\nu_n(x_1, \cdot), x_1 \in \Omega_1\}$ satisfies the LDP continuously in x_1 with rate function $J(x_1, x_2)$. Then the sequence of joint distributions $\{\mu_n\}$ obeys the LDP with rate function $I(x_1, x_2) = I_1(x_1) + J(x_1, x_2)$. There are several interesting applications of this

theorem in statistics. In particular the theorem shows that the joint distribution of the ordinary empirical measure of a sample and the corresponding bootstrap empirical measure obeys the LDP in the weak topology. Other applications include establishing the LDP property for several sampling distributions that arise naturally in statistics.

2.2 Bahadur slopes of tests in contaminated models.

One of the most basic problems in statistics is to test a null hypothesis concerning the location parameter assuming that we have a random sample of n observations from a normal population. Several test statistics are candidates for this testing problem: the mean test, the t test, the sign test and the Wilcoxon test. Among these test statistics, it is well known that the t-test is uniformly most powerful unbiased test if the normality assumption holds. But it is not clear that the t test will continue to be most powerful if there is a departure from normality. To study the robustness of these tests, it has been a standard practice to examine the performance of these tests under the Tukey model of contaminated alternatives. Under the Tukey model the sample consists of i.i.d. observations from the density

$$f(x) = (1 - \epsilon)\, \phi(x;\, \theta,\, 1) + \epsilon\, \phi(x;\, \theta,\, \sigma).$$

Here $\phi(x;\, \theta,\, \sigma)$ denotes the probability density function of a normal random variable with mean θ and standard deviation σ. And ϵ is a number between 0 and 1 representing the proportion of contamination. Two measures which are commonly used to compare the large sample properties of these tests are the Pitman efficiencies and the Bahadur slopes. Several authors have examined the robustness of the aforementioned test statistics, by computing the Pitman efficiencies. But not much work was done as regards to the computation of Bahadur slopes and efficiencies, since it is much harder problem.

The problem of deriving the Bahadur slopes is not an easy task and depends heavily on the theory of large deviations. In fact the problem of calculating the Bahadur slopes of test statistics provided the impetus for the development of large deviation theory. Both

the establishment of the LDP for a sequence of distributions and the identification of the rate function are essential for explicit calculations of Bahadur slopes. As an important application of the large deviation result described in Section 2.1, we have obtained the Bahadur slopes of the four test statistics in the Tukey model. From an examination of these slopes, it appears that the Wilcoxon test is the best performer in a neighborhood of the null hypothesis, even under the presence of moderate contamination, but is not the best performer uniformly over the whole region of the alternative hypothesis.

2.3 Quasi-least squares.

The statistical analysis of longitudinal discrete and continuous data has become an active research topic in recent years. Several books on the topic have also been published. Such data naturally occur when repeated observations are taken on individuals, or the data is taken on clusters or groups of subjects sharing similar characteristics. In a seminal paper, Liang and Zeger (1986, *Biometrika*, **73**, 13-22) introduced the generalized estimating equations (GEE) for analyzing longitudinal data. The main idea of Liang and Zeger (1986) is to model the dependence among the repeated measurements on each subject in the form of a "working correlation matrix" which is assumed to be a function of a vector α of parameters. An estimate of α is obtained using the Pearsonian residuals. The GEE method has become so popular that the 1986 article of Liang and Zeger has been included in Volume 3 of "Breakthroughs in Statistics." But recently Crowder (1995, *Biometrika*, **82**, 407-410) has pointed out some pitfalls with the estimation of the correlation parameters in the GEE method. First the estimate of α based on the Pearsonian residuals may not fall within the set of feasible values, leading to a complete breakdown of the estimation procedure. Second, even if it is feasible, it may not be consistent and it is subject to an uncertainty of definition which can lead to loss of efficiency of the regression parameter estimate. Furthermore, there can be no general asymptotic theory supporting existence or consistency of the joint distribution of the regression and the correlation parameter estimates.

In this research project we discovered a new approach for estimating the correlation parameter which overcomes all of the above pitfalls. We call this new approach as the Quasi-least squares (QLS) method. Not only does the QLS method yields feasible estimate for the correlation parameter α, it has several other advantages. For some commonly employed working correlation structures we have closed form solutions for the estimate of the correlation parameters. When the correlation matrix is unstructured, the QLS estimate of the correlation matrix involves a new factorization of a positive definite matrix. This factorization does not have a closed form solution, and in this research we have developed a recursive algorithm to obtain the factorization. Unlike the GEE method the QLS method can accommodate a wide range of correlation structures that are useful to analyze unbalanced and unequally spaced longitudinal data. While the QLS estimate of the regression parameter is consistent and asymptotically normal, the estimate of the correlation parameter is asymptotically biased. In this research we also obtained a modified QLS estimate of the correlation parameter which is consistent and asymptotically normal. For the structured correlation matrices the modified estimate is not only consistent but also robust among the popular working correlation structures. We have also developed some extensions of our results to analyzing multivariate repeated measurements.

3 List of all publications.

3.1 Papers published.

1. Large deviations for joint distributions and statistical applications. *Sankhyá*, Ser A., **2**, 147-166, 1997.

2. Bahadur slope of the t-statistic for a contaminated normal. (with J. Sethuraman). *Statistics and Probability Letters*, **34**, 245-250, 1997.

3. The large deviation principle for common statistical tests against a contaminated normal. (with J. Sethuraman). *Advances in Statistical Decision Theory and Methodology*, 239-252, 1997.

4. An alternative approach to the analysis of longitudinal data via generalized estimating equations. *Journal of Statistical Planning and Inference*, **63**, 39-54, 1997.

5. Analysis of serially correlated data using quasi-least squares. (with J. Shults). *Biometrics*, **54**, 1622-1630, 1998.

6. On inequalities for outlier detection in statistical data analysis. (with A. K. Vaish). *Journal of Applied Statistical Science*, **6**, 235-243, 1997.

7. On eliminating the asymptotic bias in quasi-least squares estimate of the correlation parameter. (with J. Shults). *Journal of Statistical Planning and Inference*, **76**, 145-161, 1999.

3.2 Technical reports.

1. Loss in efficiency due to misspecification of the correlation structure in GEE.

2. Analysis of growth curve model using quasi-least squares.

3.3 Papers submitted for publication.

1. Analysis of multivariate longitudinal data using quasi-least squares. (with D. Naik). Submitted to the *Journal of Statistical Planning and Inference.*

4 Degrees awarded.

1. Justine Shults, Phd 1996. Thesis title: "The analysis of unbalanced and unequally spaced longitudinal data using quasi-least squares."

5 Presentations and invited talks.

1. Large deviations and statistical applications. Statistics Colloquium, Division of Statistics, University of Virginia, Charlottesville, VA. March 7, 1997.

2. Spectral value decomposition of a positive definite matrix. Sixth International Workshop on "Matrix methods for statistics," Istanbul, Turkey, August 16-17, 1997.

3. Loss in efficiency due to misspecification of the correlation structure in GEE. 51st session of the International Statistical Institute, Istanbul, Turkey, August 18-25, 1997.

4. On eliminating the asymptotic bias in quasi-least squares estimate of the correlation parameter. International conference on recent advances in statistics and probability, ISI, Calcutta, December 29, 1997- January 1, 1998.

5. The analysis of longitudinal data using quasi-least squares. Seventh ILAS conference. Madison, Wisconsin, June 3-6, 1998.

6. The analysis of longitudinal data using quasi-least squares. IISA meeting, Hamilton, Canada, October 9-10, 1998.

7. Analysis of multivariate longitudinal data. Seventh International workshop on matrices and statistics, Fort Lauderdale, Florida, December 11-14, 1998.

8. Applications of large deviation theory to statistics. Colloquium talk. Dept. of Mathematics and Statistics, Carleton University, Ottawa, Canada. November 6, 1998.

9. Analysis of mixture models using quasi-least squares. Poster presentation at the NSF-CBMS lecture series on Generalized linear mixed models. Department of Statistics, University of Florida, Gainesville, FL, June 1999.

10. Analysis of growth curve model using quasi-least squares. 52nd Session of the International Statistical Institute, Helsinki, Finland. August 1999.

11. Statistical analysis of some multivariate models using quasi-least squares. Eighth International workshop on matrices and statistics. University of Tampere, Tampere, Finland. August 14, 1999.

6 Participating scientific personnel.

Narasinga Rao Chaganty, Principal Investigator.

Justine Shults, Graduate Student.

Sankhyā: The Indian Journal of Statistics
1997, Volume 59, Series A, Pt. 2, pp. 147-166

LARGE DEVIATIONS FOR JOINT DISTRIBUTIONS AND STATISTICAL APPLICATIONS*

By NARASINGA R. CHAGANTY
Old Dominion University, Norfolk

SUMMARY. We obtain the large deviation principle (LDP) of a sequence of probability measures $\{\mu_n\}$ on a product space $\Omega_1 \times \Omega_2$ when the corresponding sequences of marginal and conditional distributions possess LDP's. This is the large deviation analogue of the results of Sethuraman [*Sankhyā* A 23 1961, 379-386] for weak convergence. Our large deviation result for probability measures on product spaces re-establishes the main theorem of Dinwoodie and Zabell [*Ann. Probab.* 20 1992, 1147-1166] as a simple consequence, and also generalizes the LDP for product measures in Lynch and Sethuraman [*Ann. Probab.* 15 1987, 610-627]. Our main theorem is useful to establish the LDP for several statistical distributions. For example we show that, under bootstrapping, the ordinary empirical measure of a sample and the corresponding bootstrap empirical measure, jointly possess the LDP in the weak topology. Other applications include the LDP for noncentral t-distributions and parametric bootstrap methods.

1. Introduction

Let Ω be a Polish space, that is, a complete separable metric space and $\mathcal{B}$ be the Borel σ-field on Ω containing all the open and closed subsets of Ω. A function $I(x) : \Omega \to [0, \infty]$ is said to be a *rate function* if it is lower semi-continuous. Let $\{\mu_n\}$ be a sequence of probability measures on $(\Omega, \mathcal{B})$. We say that $\{\mu_n\}$ obeys the *weak large deviation principle* (WLDP) with rate function $I(x)$ (see for e.g., Lynch and Sethuraman (1987), Deuschel and Stroock (1989)) if the following conditions are satisfied:

Paper received. July 1996; revised April 1997.
AMS (1991) *subject classifications.* Primary 60F10, 62G09, 62F05, 62G30.
Key words and phrases. Large deviations, empirical measure, bootstrap, Bahadur slope, noncentral-t, Kullback-Leibler number.
• Research partially supported by the U. S. Army research office grant numbers DAAL03-91-G-0179, DAAH04-96-1-0070.

$$(1)\quad \limsup_n \frac{1}{n}\log \mu_n(K) \le -I(K) \qquad \ldots(1.1)$$

$$(2)\quad \liminf_n \frac{1}{n}\log \mu_n(G) \ge -I(G) \qquad \ldots(1.2)$$

for all compact sets K and for all open sets G of Ω. When $\{\mu_n\}$ satisfies (1.2) and also satisfies condition (1.1) for all closed sets C, we say that it obeys the *large deviation principle* (LDP). It is clear that if $\{\mu_n\}$ satisfies the LDP then it also satisfies the WLDP. The rate function $I(x)$ is known as a *proper rate function* if for each $L \ge 0$, the level set $\{x : I(x) \le L\}$ is a compact subset of Ω. Note that proper rate functions are also rate functions, since a nonnegative function is lower semi-continuous if and only if the level sets are closed.

The following definition of large deviation tightness, extensively used in large deviation theory, is useful when describing the parallels between weak convergence and the LDP, and also in simplifying several proofs. Our definition of large deviation tightness is same as the definition of exponential tightness in Dembo and Zeitouni (1993).

DEFINITION. A sequence of measures $\{\mu_n\}$ is *large deviation tight* (LD *tight*) if for each $N < \infty$, there exists a compact set K_N such that

$$\limsup_n \frac{1}{n}\log \mu_n(K_N^c) \le -N. \qquad \ldots(1.3)$$

Let $(\Omega_1, \mathcal{B}_1)$, $(\Omega_2, \mathcal{B}_2)$ be two Polish spaces with their associated Borel σ-fields. Let $\{\mu_{1n}\}$ be a sequence of probability measures on $(\Omega_1, \mathcal{B}_1)$ and $\{\nu_n(x_1, B_2)\}$ be a sequence of transition functions on $\Omega_1 \times \mathcal{B}_2$. Consider a sequence of probability measures $\{\mu_n\}$ on the product space $(\Omega, \mathcal{B}) = (\Omega_1 \times \Omega_2, \mathcal{B}_1 \otimes \mathcal{B}_2)$ given by

$$\mu_n(B_1 \times B_2) = \int_{B_1} \nu_n(x_1, B_2)\, d\mu_{1n}(x_1) \qquad \ldots(1.4)$$

for $B_i \in \mathcal{B}_i$, $i = 1, 2$.

We say that the sequence of probability transition functions $\{\nu_n(x_1, \cdot), x_1 \in \Omega_1\}$ satisfies the LDP *continuously in* x_1 with rate function $J(x_1, x_2)$, or simply the LDP *continuity condition* holds, if

(i) For each $x_1 \in \Omega_1$, $J(x_1, \cdot)$ is a proper rate function on Ω_2.

(ii) For any sequence $\{x_{1n}\}$ in Ω_1 such that $x_{1n} \to x_1$, the sequence of measures $\{\nu_n(x_{1n}, \cdot)\}$ on Ω_2 obeys the LDP with rate function $J(x_1, \cdot)$.

(iii) $J(x_1, x_2)$ is lower semi-continuous as a function of (x_1, x_2).

When (i) and (ii) alone hold, we say that the sequence of transition functions $\{\nu_n(x_1, \cdot), x_1 \in \Omega_1\}$ satisfies the *exponential continuity condition* with proper rate function $J(x_1, \cdot)$, following the definition given in (1.7) of Dinwoodie and Zabell (1992).

Suppose that the sequence $\{\mu_{1n}\}$ obeys the LDP with proper rate function $I_1(x_1)$ and the sequence of probability transition functions $\{\nu_n(x_1, \cdot), x_1 \in \Omega_1\}$ satisfies the LDP continuously in x_1 with rate function $J(x_1, x_2)$. Under these conditions, the main Theorem 2.3 of this paper shows that the sequence of joint distributions $\{\mu_n\}$ obeys the WLDP with rate function $I(x_1, x_2) = I_1(x_1) + J(x_1, x_2)$. And the sequence of marginal distributions $\{\mu_{2n}(B_2) = \mu_n(\Omega_1 \times B_2)\}$ on Ω_2 obeys the LDP with rate function $I_2(x_2) = \inf_{x_1 \in \Omega_1} [I(x_1, x_2)]$. Furthermore, the sequence $\{\mu_n\}$ obeys the LDP if $I(x_1, x_2)$ is a proper rate function. The proof uses Varadhan's theorem on asymptotic behavior of certain integrals. Theorem 2.3 generalizes Corollary 2.9 of Lynch and Sethuraman (1987) for product measures. The main theorem of Dinwoodie and Zabell (1992) also follows from our Theorem 2.3 as a special case where $\mu_{1n} = \mu$, for all n, where μ is a measure on Ω_1 with compact support.

Theorem 2.3 is useful to establish the LDP for commonly occurring statistical sampling distributions. Both the establishment of the LDP for a sequence of distributions and the identification of the rate function are essential for explicit calculations of Bahadur slopes and Chernoff indices. With this in view we illustrate the usefulness of our theorem by obtaining the LDP for the noncentral t-statistic and identify its rate function. As another application, we will use our theorem to show that the joint distribution of the ordinary empirical measure of a sample and the corresponding bootstrap empirical measure obeys the LDP in the weak topology. We also give explicit form for the rate function in terms of the Kullback-Leibler number. The preceding applications are not covered neither by the results of Lynch and Sethuraman (1987) nor by the results of Dinwoodie and Zabell (1992). Other important applications of our main theorem to establish the LDP for sample path processes can be found in Section 3.4 of Zajic (1993). Our theorem is also useful to establish the LDP for the t-statistic based on a random sample from a contaminated normal distribution, see Chaganty and Sethuraman (1997a, 1997b).

There is a long history of parallels between weak convergence and the LDP. Section 2 of Lynch and Sethuraman (1987) showed systematically the parallels between several results in weak convergence and in large deviations. This was followed by summary table in Vervaat (1988) listing further parallels. Recently, Puhalskii (1991) has carried the parallelism to Prokhorov's criterion by showing the equivalence of large deviation tightness and the existence of subsequences possessing the LDP. Our main theorem shows that the parallelism extends to the theorems in Sethuraman (1961) for weak convergence of joint distributions in terms of the convergence of marginals and conditional distributions. Some results similar to ours in the context of capacities can be found in Gerritse (1995).

The organization of this paper is as follows: In Section 2 we prove the main

theorem of this paper after stating some known theorems in large deviation theory. In Section 3, we show that a new interpretation of a theorem of Ellis (1984), provides sufficient conditions for the LDP continuity condition to hold for a sequence of probability transition functions defined on an Euclidean space. We also present examples of statistical distributions where the LDP continuity condition is satisfied. In Section 4, we establish the LDP for some sampling distributions that arise in statistical theory.

2. Preliminaries and Main Result

In this section we state and prove the main theorem of this paper. We will first present some known results in large deviations which are needed in the proof of our main theorem. The interrelationship between the LDP, WLDP and LD *tightness* is given in the following lemma, a proof of which can be found in Lynch and Sethuraman (1987), Dembo and Zeitouni (1993).

LEMMA 2.1. *Let $\{\mu_n\}$ be a sequence of probability measures defined on a Polish space Ω. Then the following hold:*

(1) *If $\{\mu_n\}$ is LD tight and obeys the WLDP with rate function $I(x)$ then $I(x)$ is a proper rate function and $\{\mu_n\}$ obeys the LDP with proper rate function $I(x)$.*

(2) *If $\{\mu_n\}$ obeys the LDP with proper rate function $I(x)$, then $\{\mu_n\}$ is LD tight.*

The following theorem due to Varadhan (1966), plays an important role in the proof of our main theorem. The special case of Theorem 2.2, where $F_n = F$, $\forall n$, and F is a bounded continuous function, is widely quoted in large deviation theory, and is known as Varadhan's theorem on the asymptotics of integrals. See Ellis (1985), Lynch and Sethuraman (1987). Theorem 2.2 is simply a combination of Theorems 3.2, 3.3 and 3.5 in Varadhan (1966), and is useful in the proof of our main theorem.

THEOREM 2.2 (Varadhan). *Let Ω_1 be a Polish space. Let $\{\mu_{1n}\}$ be a sequence of probability measures on $(\Omega_1, \mathcal{B}_1)$. Assume that $\{\mu_{1n}\}$ obeys the LDP with proper rate function $I_1(x_1)$. Let $\{F_n(x_1)\}$ be a sequence of real valued functions and $F(x_1)$ be another real valued function. Let*

$$\Xi_n(B_1) = \int_{B_1} exp(nF_n(x_1))\, d\mu_{1n}(x_1) \qquad \ldots(2.1)$$

for $B_1 \in \mathcal{B}_1$. Then the following hold:

(1) *Assume that there exists a constant* $L < \infty$ *such that* $F_n(x_1) \le L$ *for all* n, $x_1 \in \Omega_1$. *Suppose that* $\limsup_n F_n(x_{1n}) \le F(x_1)$ *for any sequence* $x_{1n} \to x_1$. *Then*

$$\limsup_n \frac{1}{n} \log \Xi_n(C_1) \le \sup_{x_1 \in C_1} [F(x_1) - I(x_1)] \qquad \ldots(2.2)$$

for any closed subset C_1 *of* Ω_1.

(2) *Suppose that* $\liminf_n F_n(x_{1n}) \ge F(x_1)$ *for any sequence* $x_{1n} \to x_1$. *Then*

$$\liminf_n \frac{1}{n} \log \Xi_n(G_1) \ge \sup_{x_1 \in G_1} [F(x_1) - I(x_1)] \qquad \ldots(2.3)$$

for any open set G_1 *of* Ω_1.

We now state the main theorem of this paper.

THEOREM 2.3. *Let* $(\Omega_1, \mathcal{B}_1)$, $(\Omega_2, \mathcal{B}_2)$ *be two Polish spaces with their associated Borel* σ*-fields. Let* $\{\mu_{1n}\}$ *be a sequence of probability measures on* $(\Omega_1, \mathcal{B}_1)$. *Let* $\{\nu_n(x_1, B_2)\}$ *be a sequence of probability transition functions defined on* $\Omega_1 \times \mathcal{B}_2$. *Suppose that the following two conditions are satisfied:*

(a) $\{\mu_{1n}\}$ *obeys the LDP with proper rate function* $I_1(x_1)$.

(b) $\{\nu_n(x_1, B_2)\}$ *obeys the LDP continuity condition with rate function* $J(x_1, x_2)$.

Then the sequence of joint distributions $\{\mu_n\}$ *given by (1.4) on the product space* $\Omega = \Omega_1 \times \Omega_2$, *obeys the WLDP with rate function* $I(x_1, x_2) = I_1(x_1) + J(x_1, x_2)$. *And the sequence of marginal distributions* $\{\mu_{2n}\}$ *obeys the LDP with rate function* $I_2(x_2)$. *Moreover,* $\{\mu_n\}$ *satisfies the LDP if* $I(x_1, x_2)$ *is a proper rate function.*

We will first prove a simple lemma.

LEMMA 2.4. *Let* $I_1(x_1)$ *be a proper rate function on* Ω_1 *and* $J(x_1, x_2)$ *be a rate function on* $\Omega_1 \times \Omega_2$. *Then*

$$I_2(x_2) = \inf_{x_1 \in \Omega_1} [I_1(x_1) + J(x_1, x_2)] \qquad \ldots(2.4)$$

is a rate function on Ω_2.

PROOF. Let $L \ge 0$ be fixed. It suffices to show that the set $M = \{x_2 : I_2(x_2) \le L\}$ is closed. Let $\{x_{2n}\} \in M$ be such that $x_{2n} \to x_2^*$. Choose a sequence $\{x_{1n}\}$ such that

$$I_1(x_{1n}) + J(x_{1n}, x_{2n}) \le I_2(x_{2n}) + \frac{1}{n} \le L + \frac{1}{n} \qquad \forall\, n. \qquad \ldots(2.5)$$

Since $J(x_1, x_2)$ is a nonnegative function, (2.5) implies that $I_1(x_{1n}) \leq L + 1$ for all $n \geq 1$. But the set $\{x_1 : I_1(x_1) \leq L + 1\}$ is compact, and therefore there exists a subsequence $\{x^*_{1n}\}$ of $\{x_{1n}\}$ such that $x^*_{1n} \to x^*_1$ as $n \to \infty$. Since $J(x_1, x_2)$ is lower semi-continuous in (x_1, x_2), from (2.5) it follows that

$$\begin{aligned} I_2(x^*_2) &\leq I_1(x^*_1) + J(x^*_1, x^*_2) \\ &\leq \liminf_n I_1(x^*_{1n}) + \liminf_n J(x^*_{1n}, x^*_{2n}) \qquad \ldots(2.6) \\ &\leq L. \end{aligned}$$

Thus $x^*_2 \in M$. This completes the proof of the lemma. □

PROOF OF THEOREM 2.3. We first note that $I(x_1, x_2)$ is a rate function on Ω. We will need this fact below in (2.10). Using Varadhan's theorem we will establish the upper bound (1.1) for closed rectangular sets. Let C_1 and C_2 be closed subsets of Ω_1 and Ω_2 respectively. If $F_n(x_1) = \frac{1}{n} \log [\nu_n(x_1, C_2)]$ then

$$\begin{aligned} \mu_n(C_1 \times C_2) &= \int_{C_1} \nu_n(x_1, C_2)\, d\mu_{1n}(x_1) \\ &= \int_{C_1} \exp(n F_n(x_1))\, d\mu_{1n}(x_1). \end{aligned} \qquad \ldots(2.7)$$

Note that $F_n(x_1) \leq 0$ and $\limsup_n F_n(x_{1n}) \leq -J(x_1, C_2)$ whenever $x_{1n} \to x_1$. Thus by Theorem 2.2 *(1)*, we get

$$\begin{aligned} \limsup_n \frac{1}{n} \log \mu_n(C_1 \times C_2) &= \limsup_n \frac{1}{n} \log \int_{C_1} \exp(nF_n(x_1))\, d\mu_{1n}(x_1) \\ &\leq -\inf_{x_1 \in C_1} [I_1(x_1) + J(x_1, C_2)] \\ &= -I(C_1 \times C_2) \end{aligned} \qquad \ldots(2.8)$$

and therefore the upper bound (1.1) for closed rectangular sets. In particular choosing $C_1 = \Omega_1$ in (2.8) we get

$$\limsup_n \frac{1}{n} \log \mu_{2n}(C_2) \leq -I(\Omega_1 \times C_2) = -I_2(C_2). \qquad \ldots(2.9)$$

Let $K \subset \Omega = \Omega_1 \times \Omega_2$ be compact and $l < I(K)$. For each $(x_1, x_2) \in K$, since $I(\cdot)$ is lower semi-continuous, there are open sets $O^i_{x_i}$ in Ω_i containing x_i, $i = 1, 2$ such that

$$I(O^1_{x_1} \times O^2_{x_2}) = \inf\{I(y_1, y_2) : (y_1, y_2) \in O^1_{x_1} \times O^2_{x_2}\} > l \qquad \ldots(2.10)$$

Furthermore, since Ω_i is Polish, we can find open subsets $N^i_{x_i}$ of $O^i_{x_i}$ such that $x_i \in N^i_{x_i}$ and $\overline{N}^i_{x_i} \subset O^i_{x_i}$. Consider the open covering $\cup_{(x_1,x_2)\in K} N^1_{x_1} \times N^2_{x_2}$ of K. Because K is compact we can extract a finite subcovering $\cup_{j=1}^m N^1_{x_{1j}} \times N^2_{x_{2j}}$ for K. Since $\overline{N}^i_{x_{ij}}$ is closed and K is a subset of $\cup_{j=1}^m \overline{N}^1_{x_{1j}} \times \overline{N}^2_{x_{2j}}$ we get

$$\begin{aligned}
\limsup_n \frac{1}{n} \log \mu_n(K) &\leq \max_{1\leq j\leq m} \limsup_n \frac{1}{n} \log \mu_n(\overline{N}^1_{x_{1j}} \times \overline{N}^2_{x_{2j}}) \\
&\leq -\min_{1\leq j\leq m} \{I(\overline{N}^1_{x_{1j}} \times \overline{N}^2_{x_{2j}})\} \qquad \ldots(2.11) \\
&\leq -\min_{1\leq j\leq m} \{I(O^1_{x_{1j}} \times O^2_{x_{2j}})\} \\
&\leq -l,
\end{aligned}$$

for each $l < I(K)$, and hence $\limsup_n n^{-1} \log \mu_n(K) \leq -I(K)$. This shows that $\{\mu_n\}$ satisfies (1.1) for any compact set K. We proceed similarly to obtain the lower bound. First using Varadhan's theorem we show that the lower bound (1.2) holds for any open rectangular set and then deduce (1.2) for any open set. Now let G_i be open in Ω_i for $i = 1, 2$. We can write

$$\begin{aligned}
\mu_n(G_1 \times G_2) &= \int_{G_1} \nu_n(x_1, G_2)\, d\mu_{1n}(x_1) \qquad \ldots(2.12) \\
&= \int_{G_1} \exp(n F_n(x_1))\, d\mu_{1n}(x_1)
\end{aligned}$$

where $F_n(x_1) = \frac{1}{n} \log[\nu_n(x_1, G_2)]$. Since $\liminf_n F_n(x_{1n}) \geq F(x_1) = -J(x_1, G_2)$ for any sequence $x_{1n} \to x_1$, using Theorem 2.2 *(2)*, we get

$$\begin{aligned}
\liminf_n \frac{1}{n} \log \mu_n(G_1 \times G_2) &= \liminf_n \frac{1}{n} \log \int_{G_1} \exp(nF_n(x_1))\, d\mu_{1n}(x_1) \\
&\geq -\inf_{x_1\in G_1} [I_1(x_1) + J(x_1, G_2)] \\
&= -I(G_1 \times G_2). \qquad \ldots(2.13)
\end{aligned}$$

Choosing $G_1 = \Omega_1$ in (2.13) we get

$$\liminf_n \frac{1}{n} \log \mu_{2n}(G_2) \geq -I(\Omega_1 \times G_2) = -I_2(G_2). \qquad \ldots(2.14)$$

Now let G be an open set in $\Omega_1 \times \Omega_2$. Fix $\epsilon > 0$ and choose (x_1, x_2) satisfying $I(x_1, x_2) < I(G) + \epsilon$. There exist open sets O_{x_i} in Ω_i containing x_i, $i = 1, 2$ such that $O_{x_1} \times O_{x_2} \subset G$. Thus

$$\begin{aligned} \liminf_n \frac{1}{n} \log \mu_n(G) &\geq \liminf_n \frac{1}{n} \log \mu_n(O_{x_1} \times O_{x_2}) \\ &\geq -I(x_1, x_2) \\ &\geq -I(G) - \epsilon. \end{aligned} \qquad \ldots(2.15)$$

Since $\epsilon > 0$ is arbitrary, this establishes (1.2) for any open set G. Thus $\{\mu_n\}$ obeys the WLDP. From (2.9) and (2.14) and Lemma 2.4 we can see that $\{\mu_{2n}\}$ obeys the LDP with rate function $I_2(x_2)$.

Let us now assume that $I(x_1, x_2)$ is a proper rate function. By the Contraction principle (see Appendix), $I_2(x_2)$ is also a proper rate function. Therefore $\{\mu_{2n}\}$ obeys the LDP with proper rate function $I_2(x_2)$ and hence it is LD tight. By Lemma 2.1 *(1)*, the last assertion of Theorem 2.3 will be established if we show that $\{\mu_n\}$ is LD tight. Since $\{\mu_{in}\}$ is LD tight, given $N < \infty$, we can find a compact subset K_i of Ω_i such that

$$\limsup_n \frac{1}{n} \log \mu_{in}(K_i^c) \leq -2N, \qquad \ldots(2.16)$$

for $i = 1, 2$. Hence there exists n_0 such that

$$\mu_{in}(K_i^c) \leq \exp(-nN) \qquad \ldots(2.17)$$

for $i = 1, 2$ and $n \geq n_0$. Let $K = K_1 \times K_2$, which is compact in $\Omega_1 \times \Omega_2$, being the product of two compact sets. For $n \geq n_0$ from (2.17) we get

$$\begin{aligned} \mu_n(K^c) &\leq \mu_{1n}(K_1^c) + \mu_{2n}(K_2^c) \\ &\leq 2 \exp(-nN) \end{aligned} \qquad \ldots(2.18)$$

which implies that

$$\limsup_n \frac{1}{n} \log \mu_n(K^c) \leq -N. \qquad \ldots(2.19)$$

This completes the proof of Theorem 2.3. □

REMARK 2.5. It is interesting to note that $\{\mu_{2n}\}$ satisfies the upper bound (2.9) for closed sets and the lower bound (2.14) for open sets, even if $J(x_1, x_2)$ is not lower semi-continuous in (x_1, x_2). The lower semi-continuity of $J(x_1, x_2)$ as a function of (x_1, x_2) is needed to show that $I_2(x_2)$ is a rate function.

We shall now give some sufficient conditions on the rate functions $I_1(x_1)$ and $J(x_1, x_2)$ which guarantee that $I(x_1, x_2)$ is a proper rate function.

LEMMA 2.6. *Let Ω_1 and Ω_2 be two Polish spaces. Let $J : \Omega = \Omega_1 \times \Omega_2 \to [0, \infty]$ be a function satisfying the following conditions:*

(a) *The function* $J(x_1, x_2)$ *is lower semi-continuous in* (x_1, x_2).

(b) *The set* $\cup_{x_1 \in K_1} \{x_2 : J(x_1, x_2) \leq L\}$ *is a compact subset of* Ω_2 *for any* $L \geq 0$ *and for any compact set* K_1 *of* Ω_1.

Let $I_1(x_1)$ *be a proper rate function defined on* Ω_1 *and let* $I(x_1, x_2) = I_1(x_1) + J(x_1, x_2)$. *Then the following hold:*

(1) For each $x_1 \in \Omega_1$, $J(x_1, \cdot)$ *is a proper rate function on* Ω_2.

(2) $I(x_1, x_2)$ *is a proper rate function on* Ω.

PROOF. It is easy to see that conditions (a) and (b) imply that $J(x_1, \cdot)$ is a proper rate function on Ω_2 and $I(x_1, x_2)$ is lower semi-continuous. We proceed to show that $I(x_1, x_2)$ has compact level sets. Fix $L \geq 0$ and let

$$M = \{(x_1, x_2) : I(x_1, x_2) \leq L\}. \qquad \ldots(2.20)$$

Note that M is a closed subset of Ω since $I(x_1, x_2)$ is lower semi-continuous. Let $K_1 = \{x_1 : I_1(x_1) \leq L\}$. It is easy to verify that

$$M \subset K_1 \times \bigcup_{x_1 \in K_1} \{x_2 : J(x_1, x_2) \leq L\}. \qquad \ldots(2.21)$$

Since K_1 is compact, the set on the right hand side is compact by condition (b). Thus M is compact being a closed subset of a compact set. This completes the proof of the Lemma 2.6. □

Suppose now that Ω_1 be a locally compact, separable metric space and $J(\cdot, x_2)$ is lower semi-continuous on Ω_1 for each $x_2 \in \Omega_2$. Lemma 2.7 below shows that if $J(x_1, x_2)$ satisfies condition (b) of Lemma 2.6, then it is lower semi-continuous as a function of (x_1, x_2). Thus in this case where Ω_1 is R^d, we need only verify that $J(x_1, x_2)$ is lower semi-continuous in x_1 for fixed x_2, and satisfies condition (b) to conclude that $I(x_1, x_2)$ is a proper rate function.

LEMMA 2.7. *Let* Ω_1 *be a locally compact, separable metric space and* Ω_2 *be a Polish space. Let* $J(x_1, x_2)$ *be a nonnegative function on the product space* $\Omega = \Omega_1 \times \Omega_2$ *such that* $J(\cdot, x_2)$ *is lower semi-continuous on* Ω_1 *for each* $x_2 \in \Omega_2$ *and condition* (b) *of Lemma 2.6 holds. Then* $J(x_1, x_2)$ *is lower semi-continuous on* Ω.

PROOF. We first note that Ω_1 is a Polish space, since a locally compact metric space is topologically complete (see Dugundji (1970); Corollary 2.4 page 294). Let $L \geq 0$. It suffices to show that $M = \{(x_1, x_2) : J(x_1, x_2) \leq L\}$ is closed in Ω. Note that we can represent the set M as

$$M = \bigcup_{x_1 \in \Omega_1} (\{x_1\} \times M_{x_1}) \qquad \ldots(2.22)$$

where $M_{x_1} = \{x_2 : J(x_1, x_2) \leq L\}$, for $x_1 \in \Omega_1$. Let (x_{1n}, x_{2n}) be a sequence of points in M such that $(x_{1n}, x_{2n}) \to (x_1^*, x_2^*)$ as $n \to \infty$. Since Ω_1 is locally compact there exists $K_{11} \supset K_{12} \cdots$, a countable, compact neighborhood base of x_1^*. Now for each $j \geq 1$, we can find n_j such that $x_{1n} \in K_{1j}$ for all $n \geq n_j$. Thus we have for $n \geq n_j$,

$$x_{2n} \in \bigcup_{x_1 \in K_{1j}} M_{x_1} \qquad \ldots(2.23)$$

which is a compact set by hypothesis **(b)**. Hence $x_2^* \in \cup_{x_1 \in K_{1j}} M_{x_1}$ for all $j \geq 1$. Since $J(\cdot, x_2)$ is lower semi-continuous in the first coordinate we can verify that

$$\bigcap_{j=1}^{\infty} \bigcup_{x_1 \in K_{1j}} M_{x_1} = M_{x_1^*} \qquad \ldots(2.24)$$

which implies that $x_2^* \in M_{x_1^*}$, that is, $(x_1^*, x_2^*) \in M$. This completes the proof of the lemma. □

REMARK 2.8. Assume that $\{\mu_{1n}\}$ obeys the LDP with proper rate function $I_1(x_1)$. Suppose that the transition functions $\nu_n(x_1, B_2) = \mu_{2n}(B_2)$, do not depend on x_1 and $\{\mu_{2n}(B_2)\}$ obeys the LDP with proper rate function given by $I_2(x_2)$. Then it trivially follows from Theorem 2.3 that the sequence of joint distributions $\{\mu_n\}$ obeys the LDP with proper rate function

$$I(x_1, x_2) = I_1(x_1) + I_2(x_2), \qquad \ldots(2.25)$$

a result originally obtained by Lynch and Sethuraman (1987); see Corollary 2.9 in their paper.

REMARK 2.9. Theorem 2.3 of Dinwoodie and Zabell (1992) can be deduced from our Theorem 2.3 as follows : Suppose that $\mu_{1n} \equiv \mu \; \forall n$ and the support of μ, say $S(\mu)$ is compact. Then it is easy to see that the sequence $\{\mu_{1n}\}$ obeys the LDP with proper rate function

$$I_1(x_1) = \begin{cases} 0 & \text{if } x_1 \in S(\mu) \\ \infty & \text{otherwise.} \end{cases} \qquad \ldots(2.26)$$

Simply put, a single measure $\{\mu\}$ always obeys the LDP with the above rate function. Further the rate function defined in (2.26) is a proper rate function if and only if the support $S(\mu)$, of the measure μ, is compact. With this choice of $\{\mu_{1n} \equiv \mu\}$ under the conditions of Theorem 2.3 of Dinwoodie and Zabell (1992), it follows from our Theorem 2.3 that the sequence of marginal measures $\{\mu_{2n}\}$ obeys the LDP with rate function $I_2(x_2) = \inf_{x_1 \in S(\mu)} J(x_1, x_2)$.

3. LDP Continuity for Probability Transition Functions on R^d

In this section we will examine sufficient conditions for the LDP continuity condition to hold for an arbitrary sequence of probability transition functions defined on an Euclidean space. Dinwoodie and Zabell (1992) have given some sufficient conditions for exponential continuity in the case where Ω_1 is a first countable topological space, Ω_2 is a locally convex Hausdorff space and $\nu_n(x_1, B_2)$ is the distribution of the average of n i.i.d. random vectors. However their sufficient conditions are very restrictive. Even in the simplest case where Ω_2 is the real line, their sufficient conditions are not satisfied for averages of i.i.d. random variables from basic statistical distributions, see Remark 3.7 below.

We will first discuss a method of verifying the exponential continuity condition for an arbitrary sequence of transition functions, by giving a new interpretation to a theorem of Ellis (1984).

Let Ω_1 is a Polish space and for each $x_1 \in \Omega_1$, let $\{\nu_n(x_1, \cdot)\}$ be a sequence of probability transition functions on $\Omega_2 = \mathcal{R}^{d_2}$. We can verify that the exponential continuity condition holds for the sequence $\{\nu_n(x_1, \cdot), x_1 \in \Omega_1\}$, using a theorem of Ellis (1984) as follows: Let $\{x_{1n}\}$ and x_1 be in Ω_1 such that $x_{1n} \to x_1$. Let $\{Y_n\}$ be a sequence of $\mathcal{R}^{d_2}$ valued random variables such that the distribution of Y_n/n is given by $\nu_n(x_{1n}, \cdot)$. Define

$$c_n(x_{1n}, t) = \frac{1}{n} \log E\big[\exp(< t, Y_n >)\big]. \qquad \ldots(3.1)$$

Suppose that $\lim_n c_n(x_{1n}, t)$ exists and equals $c(x_1, t)$, for all $t \in \mathcal{R}^{d_2}$, where we allow $+\infty$ both as a limit value and as an element in the sequence $\{c(x_{1n}, t)\}$. Let $\mathcal{D}_{x_1}(c) = \{t \in \mathcal{R}^{d_2} : c(x_1, t) < \infty\}$. The function $c(x_1, t) : \mathcal{R}^{d_2} \to \mathcal{R}$ is said to be closed if $\{t : c(x_1, t) \leq \alpha\}$ is closed for each real α. This is equivalent to $c(x_1, t)$ being lower semi-continuous. If $c(x_1, t)$ is differentiable on the interior of $\mathcal{D}_{x_1}(c)$, then we call $c(x_1, t)$ steep if $\| \operatorname{grad}(c(x_1, t_n)) \| \to \infty$, for any sequence $\{t_n\} \subset \operatorname{int}(\mathcal{D}_{x_1}(c))$ which tends to a boundary point of $\mathcal{D}_{x_1}(c)$. For $x_2 \in \mathcal{R}^{d_2}$, let

$$J(x_1, x_2) = \sup_{t \in \mathcal{R}^{d_2}} [< t, x_2 > - c(x_1, t)], \qquad \ldots(3.2)$$

be the Legendre-Fenchel transform of $c(x_1, t)$.

THEOREM 3.1 (Ellis). *If $\mathcal{D}_{x_1}(c)$ has a nonempty interior containing the point $t=0$ and $c(x_1, t)$ is a closed, convex function of $\mathcal{R}^{d_2}$ then the function $J(x_1, \cdot)$ defined in (3.2) is a proper rate function on $\mathcal{R}^{d_2}$ and the sequence of probability measures $\{\nu_n(x_{1n}, \cdot)\}$ satisfies the upper bound (1.1) for all closed sets C of $\mathcal{R}^{d_2}$ with proper rate function $J(x_1, \cdot)$. Furthermore, if $c(x_1, t)$ is differentiable on all of interior of $\mathcal{D}_{x_1}(c)$ and is steep then the sequence of probability measures*

$\{\nu_n(x_{1n}, \cdot)\}$ satisfies the lower bound (1.2) for all open sets G of $\mathcal{R}^{d_2}$ with proper rate function $J(x_1, \cdot)$.

The next lemma due to Dinwoodie and Zabell (1992), (see Lemma 3.1 (i) in their paper), provides a simple sufficient condition for the function $J(x_1, x_2)$ defined by (3.2) to be lower semi-continuous in (x_1, x_2). Note that Theorem 3.1 in conjunction with Lemma 3.2, provides sufficient conditions for a sequence of probability transition functions on an Euclidean space, to satisfy the LDP continuity condition.

LEMMA 3.2. *Let $\{x_{1n}\}$ be a sequence and $\{x_1\}$ in Ω_1 be such that $x_{1n} \to x_1$. If for every $t \in \mathcal{R}^{d_2}$, there exits a sequence $t_n \to t$ such that*

$$\limsup_n c(x_{1n}, t_n) \leq c(x_1, t) \qquad \ldots(3.3)$$

then the function $J(x_1, x_2)$ defined by (3.2) is lower semi-continuous in (x_1, x_2).

We will now present a number of examples where the LDP continuity condition holds. In Examples 3.3, 3.4, 3.5 and 3.6 below, Y_n is the sum of n i.i.d. random variables $X_1, \ldots, X_n$. Furthermore, the common distribution η_θ of the X_i's is indexed by a parameter $\theta \in \Omega_1$. Hence the function defined by (3.1), is independent of n, but depends on the parameter θ and therefore we shall denote it by $c(\theta, t)$. In all of the four examples, it is easy to verify that $c(\theta, t)$ as a function of t is closed, convex and steep, for fixed $\theta \in \Omega_1$. Clearly, $c(\theta, t)$ is continuous in θ for each t in these examples and therefore condition (3.3) is trivially satisfied. Let the distribution of the sample mean $\overline{X}_n = Y_n/n$ be given by $\nu_n(\theta, \cdot)$. Thus it follows from Theorem 3.1 and Lemma 3.2, that the sequence of probability transition functions $\{\nu_n(\theta, \cdot), \theta \in \Omega_1\}$, in all of the four examples, satisfies the LDP continuity condition with rate function $J(\theta, z) = \sup_{t \in R}[z\,t - c(\theta, t)]$. We will omit details and present only the distribution η_θ, the function $c(\theta, t)$ and the rate function $J(\theta, z)$ in the examples below.

In Examples 3.3, 3.4, 3.5 and 3.6, the random variables are defined to be degenerate, when θ is a boundary point of Ω_1. In (3.8) and elsewhere in this paper we let $0 \log \frac{0}{0} = 0$.

EXAMPLE 3.3. Let $X_1, \ldots, X_n$ be i.i.d. normal with mean θ_1 and variance θ_2. If we let $\theta = (\theta_1, \theta_2)$ then $\theta \in \Omega_1 = (-\infty, \infty) \times [0, \infty)$. It is easy to verify that

$$c(\theta, t) = \theta_1 t + \frac{1}{2}\theta_2 t^2, \quad -\infty < t < \infty, \qquad \ldots(3.4)$$

for $\theta \in \Omega_1$. The sequence of probability distributions of the sample means $\{\overline{X}_n\}$ satisfies the LDP continuity condition with rate function

$$J(\theta, z) = \frac{(z - \theta_1)^2}{2\theta_2}, \quad -\infty < z < \infty, \qquad \ldots(3.5)$$

for $\theta_2 > 0$, $-\infty < \theta_1 < \infty$; and

$$J(\theta, z) = \begin{cases} 0 & \text{if } z = \theta_1 \\ \infty & \text{otherwise} \end{cases} \qquad \ldots(3.6)$$

when $\theta_2 = 0$ and $-\infty < \theta_1 < \infty$.

EXAMPLE 3.4. Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli with mean θ, $\theta \in \Omega_1 = [0, 1]$. The function $c(\theta, t)$ is given by

$$c(\theta, t) = \log[\theta \exp(t) + (1 - \theta)], \qquad -\infty < t < \infty. \qquad \ldots(3.7)$$

for $\theta \in \Omega_1$. The sequence of probability distributions of the sample means $\{\overline{X}_n\}$ satisfies the LDP continuity condition with rate function

$$J(\theta, z) = \begin{cases} z \log \dfrac{z}{\theta} + (1 - z) \log \dfrac{(1 - z)}{(1 - \theta)} & \text{if } 0 \le z \le 1 \\ \infty & \text{otherwise,} \end{cases} \qquad \ldots(3.8)$$

for $\theta \in \Omega_1$.

EXAMPLE 3.5. Let $X_1, \ldots, X_n$ be i.i.d. $f_\theta(x) = \dfrac{1}{\theta^\alpha \Gamma(\alpha)} \exp(-x/\theta)\, x^{(\alpha-1)}$, $x > 0$, $\alpha > 0$ and $\theta \in \Omega_1 = [0, \infty)$. Then for any $\theta \in \Omega_1$, we have

$$c(\theta, t) = \begin{cases} -\alpha \log(1 - \theta t) & \text{if } t < 1/\theta \\ \infty & \text{otherwise.} \end{cases} \qquad \ldots(3.9)$$

The sequence of probability distributions of the sample means $\{\overline{X}_n\}$ satisfies the LDP continuity condition with rate function

$$J(\theta, z) = \begin{cases} \dfrac{1}{\theta}[z + \alpha\theta(\log(\alpha\theta) - 1 - \log(z))] & \text{if } z > 0, \\ \infty & \text{otherwise.} \end{cases}$$

for each $\theta > 0$; and

$$J(0, z) = \begin{cases} 0 & \text{if } z = 0 \\ \infty & \text{otherwise.} \end{cases} \qquad \ldots(3.10)$$

In the special case where $\alpha = 1/2$ and $\theta = 2$, the distribution of $n\overline{X}_n$ is $\chi^2(n)$. Thus if Y_n is $\chi^2(n)$, then $\{Y_n/n\}$ obeys the LDP with proper rate function

$$J(z) = \begin{cases} \frac{1}{2}[z - 1 - \log(z)] & \text{if } z > 0 \\ \infty & \text{otherwise.} \end{cases} \qquad \ldots(3.11)$$

EXAMPLE 3.6 Let $X_1, \ldots, X_n$ be i.i.d. $f_\theta(x) = \theta \exp(\theta(x - \mu))$, $x \le \mu$ and $\theta \in \Omega_1 = (0, \infty)$. The function $c(\theta, t)$ is given by

$$c(\theta, t) = \begin{cases} t\mu + \log(\theta) - \log(\theta + t) & \text{if } \theta + t > 0 \\ \infty & \text{otherwise.} \end{cases} \qquad \ldots(3.12)$$

Then $\{\overline{X}_n\}$ satisfies the LDP continuity condition with rate function

$$J(\theta, z) = \begin{cases} \theta(\mu - z) - \log(\theta(\mu - z)) - 1 & \text{if } z < \mu \\ \infty & \text{otherwise,} \end{cases} \quad \ldots(3.13)$$

for $\theta > 0$.

REMARK 3.7. Dinwoodie and Zabell (1992) considered Example 3.6 restricting the parameter space Ω_1 to $[\theta_0, \infty)$ where $\theta_0 > 0$ and showed that the exponential continuity condition is satisfied for $\{\overline{X}_n\}$ using a sufficient condition, which they attribute it to de Acosta, see (3.9) and Example 3.3 of their paper. We can verify that de Acosta's condition is not satisfied for the full family in Example 3.6, even though the exponential continuity condition holds.

4. Statistical Applications

Application 4.1. *LDP for noncentral t-distributions.* In Section 3 we have seen that the LDP continuity condition is satisfied for the probability distributions of averages of i.i.d. random variables from basic statistical distributions. In this section, we will demonstrate with an example that Theorem 2.3 can be used to show that the LDP holds for other statistical distributions derived from the basic statistical distributions. More specifically in Example 1.1. we show that the noncentral t-distributions obeys the LDP and identify its rate function. The rate function for the central t-distributions was derived by several authors including Sievers (1969), Bahadur (1971), Killeen *et al* (1972) and more recently by Berk (1982) and Singh (1989) for the noncentral t-distributions using different methods.

EXAMPLE 4.1. Let $X_1, \ldots, X_n$ be i.i.d. normal with mean θ and variance 1. Let $\overline{X} = \sum_{i=1}^n X_i/n$ be the sample mean and $S_n^2 = \sum_{i=1}^n (X_i - \overline{X})^2/(n-1)$ be the sample variance. Let $\Omega_1 = (0, \infty)$ and $\Omega_2 = (-\infty, \infty)$, which are topologically complete and separable metric spaces. Let $T_n = \overline{X}/\sqrt{S_n^2}$ be the noncentral t-statistic. Note that $(n-1)S_n^2$ is distributed as χ^2 with $(n-1)$ degrees of freedom and hence S_n^2 obeys the LDP with proper rate function

$$I_1(s) = \frac{1}{2}[s - 1 - \log(s)] \quad (4.1)$$

for $s \in \Omega_1$. Let $\nu_n(s, \cdot)$ be the conditional distribution of T_n given $S_n^2 = s > 0$. Since $\nu_n(s, \cdot)$ is just the normal distribution with mean $\theta/\sqrt{s}$ and variance $1/(ns)$, by Example 3.3 we have that $\{\nu_n(s, \cdot), s \in \Omega_1\}$ satisfies the LDP continuity condition with rate function

$$J_\theta(s, t) = \frac{s(t - \theta/\sqrt{s})^2}{2}, \quad -\infty < t < \infty \quad \ldots(4.2)$$

for $s \in \Omega_1$. We can easily verify that the rate function $J_\theta(s, t)$ in (4.2) satisfies condition (b) of Lemma 2.6. Therefore it follows from Theorem 2.3 that the joint distribution of (S_n^2, T_n) satisfies the LDP with proper rate function

$$I_\theta(s, t) = I_1(s) + J_\theta(s, t), \quad s > 0, \ -\infty < t < \infty. \qquad \ldots(4.3)$$

Furthermore, the marginal distribution of T_n obeys the LDP with proper rate function

$$\begin{aligned} I_2(t) &= \inf_{s \in \Omega_1} [I_1(s) + J_\theta(s, t)] \\ &= \inf_{s>0} [(s - 1 - \log(s))/2 + (t\sqrt{s} - \theta)^2/2] \end{aligned} \qquad \ldots(4.4)$$

which after simplification reduces to

$$I_2(t) = \frac{1}{2}[\theta^2 - st\theta - 2\log(s)], \quad -\infty < t < \infty, \qquad \ldots(4.5)$$

where s is the positive root of the quadratic equation $s^2(1 + t^2) - st\theta - 1 = 0$. Thus for any measurable subset A of the real line we have

$$-I_2(A^0) \le \liminf_n \frac{1}{n} \log Pr(T_n \in A) \le \limsup_n \frac{1}{n} \log Pr(T_n \in A) \le -I_2(\overline{A}). \qquad \ldots(4.6)$$

The above inequality (4.6) was established by other authors, using complex analytical calculations, when A is an interval of the form $A = [t, \infty)$, $t > \theta$.

Application 4.2. *Bootstrap Empirical Measure.* The resampling procedure, known as the "bootstrap" introduced by Efron (1979) has become very popular in statistical methodology in recent years. A formal description of the method is as follows: Let E be a Polish space and Ω_1 be the class of all probability measures defined on the collection of all Borel subsets of E endowed with the topology of weak convergence. It is well known that (Ω_1, ρ) is a Polish space, where ρ is the Lévy- Prohorov metric, see Deuschel and Stroock (1989), page 64. Fix $P \in \Omega_1$. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a sample of n i.i.d. observations from P. Let $\widehat{P}_n$ be the empirical measure based on $\mathbf{X}$, that is, $\widehat{P}_n(B)$ is simply the proportion of X_i's, $1 \le i \le n$, with values in the Borel set B of E. Given $\mathbf{X} = \mathbf{x}$, the bootstrap method is to take a random sample of n i.i.d. observations $\mathbf{X}^* = (X_1^*, \ldots, X_n^*)$ from $\widehat{P}_n$. The bootstrap empirical measure $\widetilde{P}_n$ is defined as the empirical measure of the bootstrap sample $\mathbf{X}^*$. As an important application of Theorem 2.3, we will show that the sequence of joint distribution of the empirical measure $\widehat{P}_n$ and bootstrap empirical measure $\widetilde{P}_n$ obeys the LDP with proper rate function that depends on the Kullback-Leibler number. For Q, $P \in \Omega_1$, the Kullback-Leibler number is defined as

$$K(Q, P) = \begin{cases} \int_E q \log q \, dP & \text{if } Q \ll P \\ \infty & \text{otherwise} \end{cases} \qquad \ldots(4.7)$$

where q is the Radon-Nikodym derivative of Q with respect to P.

The next theorem is the main result of this section.

THEOREM 4.2. *Let Ω_1 be the class of all probability measures on a Polish space E endowed with topology of weak convergence. Fix $P \in \Omega_1$. Let $X_1, \ldots, X_n$ be i.i.d. P. Let $\widehat{P}_n$ be the empirical measure based on $X_1, \ldots, X_n$. Let $\widetilde{P}_n$ be the empirical measure of the bootstrap sample $X_1^*, \ldots, X_n^*$. Then the joint distribution of $(\widehat{P}_n, \widetilde{P}_n)$ obeys the LDP in the weak topology with proper rate function*

$$K((Q_1, Q_2), P) = [K(Q_2, Q_1) + K(Q_1, P)]. \qquad \ldots(4.8)$$

PROOF. We will prove the theorem by simply verifying the conditions of Theorem 2.3. Let $\Omega_1 = \Omega_2$ be the class of all probability measures on E. Then Ω_1 and Ω_2 are complete separable metric spaces. By Sanov's theorem (see Theorem 3.1 in Chaganty and Karandikar (1996)), we have that $\widehat{P}_n$ obeys the LDP with proper rate function $I_1(Q_1) = K(Q_1, P)$. Let $\nu_n(Q_1, B_2) = Pr(\widetilde{P}_n \in B_2 | \widehat{P}_n = Q_1)$. It follows from Theorem 3.1 and Theorem 2.5, both in Chaganty and Karandikar (1996), that the sequence of transition functions $\{\nu_n(Q_1, B_2)\}$ satisfy the LDP continuity condition with rate function $J(Q_1, Q_2) = K(Q_2, Q_1)$ where $Q_i \in \Omega_i$, $i = 1, 2$. Therefore using Lemmas 4.3, 2.6 and Theorem 2.3 we can conclude that the sequence of joint distributions of $(\widehat{P}_n, \widetilde{P}_n)$ obeys the LDP with proper rate function

$$\begin{aligned} K((Q_1, Q_2), P) &= [I_1(Q_1) + J(Q_1, Q_2)] \\ &= [K(Q_1, P) + K(Q_2, Q_1)] \\ &= [K(Q_2, Q_1) + K(Q_1, P)]. \end{aligned} \qquad \ldots(4.9)$$

This completes the proof of the theorem. □

The next lemma shows that $K(Q, P)$ satisfies condition (b) of Lemma 2.6, thus establishing that the function (4.9) is indeed a proper rate function. The proof of the following Lemma 4.3 is similar to the proof of Lemma 2.3 (a) in Groeneboom *et al* (1979).

LEMMA 4.3. *Let M_1 be a compact set of Ω_1 in the weak topology and $L \geq 0$. Then the set*

$$\mathrm{M}_2 = \{Q \in \Omega_1 : K(Q, R) \leq L \quad \text{for some} \quad R \in \mathrm{M}_1\} \qquad \ldots(4.10)$$

is also compact in the weak topology.

PROOF. Let $\epsilon > 0$ be given. Let $\alpha > 0$ be such that $(L + 1/e)/\log \alpha < \epsilon/2$ and let $\delta = \epsilon/(2\alpha)$. Since M_1 is compact we can choose a compact set $V = V_\delta$ in E such that

$$R(V^c) \leq \delta \quad \text{for all} \quad R \in \mathrm{M}_1. \qquad \ldots(4.11)$$

Let $Q \in \mathrm{M}_2$ and $R \in \mathrm{M}_1$ be such that $K(Q, R) \le L$. Let g be the Radon-Nikodym derivative of Q with respect to R. Since $x \log(x) \ge -1/e$ for $x > 0$, we have

$$\int_{\{g>\alpha\}} g \log g \, dR \le L + 1/e. \qquad \ldots(4.12)$$

Now

$$\begin{aligned} Q(V^c) &= \int_{V^c \cap \{g \le \alpha\}} g \, dR + \int_{V^c \cap \{g>\alpha\}} g \, dR \\ &\le \alpha R(V^c) + \tfrac{1}{\log \alpha} \int_{\{g>\alpha\}} g \log g \, dR \qquad \ldots(4.13) \\ &\le \alpha \delta + \tfrac{L+1/e}{\log \alpha} \le \epsilon/2 + \epsilon/2 = \epsilon. \end{aligned}$$

Therefore we have

$$Q(V^c) \le \epsilon \quad \text{for all} \quad Q \in \mathrm{M}_2, \qquad \ldots(4.14)$$

which implies that M_2 is compact. This completes the proof of the lemma. □

By Theorem 2.3 we can also conclude that the marginal distribution of $\widetilde{P}_n$ satisfies the LDP with proper rate function

$$K^*(Q, P) = \inf_{R \in \Omega_1} [K(Q, R) + K(R, P)]. \qquad \ldots(4.15)$$

It is interesting to note that $K(Q, P) \ge K^*(Q, P)$ for all $Q, P \in \Omega_1$ (equality holds iff $P = Q$). Thus the rate function of the ordinary empirical measure $\widehat{P}_n$ is always greater than the rate function of the marginal distribution of the bootstrap empirical measure $\widetilde{P}_n$.

Application 4.3. *Parametric Bootstrap.* Let $X_1, \ldots, X_n$ be i.i.d. random vectors with distribution given by η_θ indexed by the parameter $\theta \in \Omega_1 \subset \mathcal{R}^d$, where Ω_1 is a Polish space. Let $\widehat{\theta}_n = T_n(X_1, \ldots, X_n)$ belonging to Ω_1 be an estimate of θ. In parametric bootstrap method, given $\widehat{\theta}_n = z_1$, the bootstrap sample is a sequence of i.i.d. observations $X_1^*, \ldots, X_n^*$ from η_{z_1}. Let $\widetilde{\theta}_n = T_n(X_1^*, \ldots, X_n^*)$ be the estimate of θ based on the bootstrap sample $(X_1^*, \ldots, X_n^*)$. Suppose that $\widehat{\theta}_n$ satisfies the LDP continuity condition with rate function $J(\theta, z_1)$, $z_1 \in \mathcal{R}^d$. Then by Theorem 2.3 it follows that the joint distribution of $(\widehat{\theta}_n, \widetilde{\theta}_n)$ satisfies the LDP with proper rate function

$$I(\theta, (z_1, z_2)) = [J(\theta, z_1) + J(z_1, z_2)]. \qquad \ldots(4.16)$$

Also, the marginal distribution of $\widetilde{\theta}_n$ satisfies the LDP with proper rate function

$$I_2(\theta, z_2) = \inf_{z \in \Omega_1} [J(\theta, z) + J(z, z_2)]. \qquad \ldots(4.17)$$

It is interesting to note that $J(\theta, z) \ge I_2(\theta, z)$ for all (θ, z), that is the rate function of the marginal distribution of $\widetilde{\theta}_n$ is always less than the rate function

of the distribution of $\widehat{\theta}_n$. We now present a couple of examples. In Examples 4.4 and 4.5 it is easy to check that for each fixed θ, the rate functions $J(\theta, z_1)$'s are continuous, convex functions of z_1 and jointly lower semi-continuous in (θ, z_1). Furthermore, $J(\theta, z_1)$'s also satisfy condition (b) of Lemma 2.6.

EXAMPLE 4.4. Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli with parameter $\theta \in \Omega_1 = [0, 1]$. Let $\widehat{\theta}_n = \overline{X}_n$ where $\overline{X}_n$ is the sample mean. Let $X_1^*, \ldots, X_n^*$ be i.i.d. Bernoulli with parameter z_1 given $\widehat{\theta}_n = z_1$. Let $\widetilde{\theta}_n = \overline{X}_n^*$ be the estimate of θ based on the bootstrap sample $(X_1^*, \ldots, X_n^*)$. It follows then from Example 3.4 and Theorem 2.3 that the joint distribution of $(\widehat{\theta}_n, \widetilde{\theta}_n)$ satisfies the LDP with proper rate function

$$I(\theta, (z_1, z_2)) = [J(\theta, z_1) + J(z_1, z_2)] \qquad \ldots(4.18)$$

where $J(\theta, z)$ is given by (3.8). The marginal distribution of $\widetilde{\theta}_n$ also satisfies the LDP with proper rate function

$$I_2(\theta, z_2) = \inf_{z \in [0,1]} [J(\theta, z) + J(z, z_2)]. \qquad \ldots(4.19)$$

EXAMPLE 4.5. Let $X_1, \ldots, X_n$ be i.i.d. exponential with mean $1/\theta$, that is, the common pdf is given by $f_\theta(x) = \theta \exp(-\theta x)$, $x > 0$, $\theta \in \Omega_1 = (0, \infty)$. Let $\widehat{\theta}_n = 1/\overline{X}_n$, where $\overline{X}_n$ is the sample mean. Given $\widehat{\theta}_n = z$, let $(X_1^*, \ldots, X_n^*)$ be i.i.d. $f_z(x)$. Let $\widetilde{\theta}_n = \overline{X}_n^*$ be the mean of the bootstrap sample. It follows then from Example 3.5 and Theorem 2.3 that the joint distribution of $(\widehat{\theta}_n, \widetilde{\theta}_n)$ obeys the LDP with proper rate function

$$I(\theta, (z_1, z_2)) = \begin{cases} [z_1(\theta + z_2) - 2 - \log(\theta z_1^2 z_2)] & \text{if } z_1 > 0, z_2 > 0, \\ \infty & \text{otherwise,} \end{cases} \qquad \ldots(4.20)$$

for $\theta > 0$. The marginal distribution of $\widetilde{\theta}_n$ also obeys the LDP with proper rate function

$$I_2(\theta, z_2) = \begin{cases} [2\log((\theta + z_2)/2) - \log(\theta z_2)] & \text{if } z_2 > 0 \\ \infty & \text{otherwise,} \end{cases} \qquad \ldots(4.21)$$

for $\theta > 0$.

Appendix

The following result, known as the contraction principle, is a very useful device to deduce the LDP for a sequence of measures induced by a continuous function from a sequence of measures which are known to obey the LDP. An extension of the contraction principle for measurable functions h can be found in Puhalskii (1991).

Contraction principle. Let $h : \Omega \to \Omega^*$ be a continuous function where Ω and Ω^* are two Polish spaces. Let I be a proper rate function on Ω. Then $I^*(y) = \inf_{\{x: h(x)=y\}} I(x)$ is a proper rate function on Ω^*. Suppose that $\{\mu_n\}$ is a sequence of probability measures on Ω which obeys the LDP with proper rate function I. Then the sequence $\{\mu_n^* = \mu_n h^{-1}\}$ obeys the LDP with proper rate function I^*.

ACKNOWLEDGEMENTS. The author is very grateful to Professor R. L. Karandikar for valuable discussions during the preparation of this manuscript and to Professor J. Sethuraman for introducing him to the theory of large deviations.

References

BAHADUR, R.R (1971). *Some Limit Theorems in Statistics.* SIAM, Philadelphia.

BERK, R.H. (1982). On an asymptotically optimal sequential test. *Scand. Journal of Stat.* 9 159-163.

CHAGANTY, N.R. and KARANDIKAR, R.L. (1996). Some properties of the Kullback-Leibler number. *Sankhyā* A 58 69-80.

CHAGANTY, N.R. and SETHURAMAN, J. (1997a). Bahadur slope of the t-statistic for a contaminated normal. To appear in *Statist. & Prob. Letters.*

---- (1997b). The large deviation principle for common statistical tests against a contaminated normal. To appear in *Advances in Statistical Decision Theory and Methodology*, Birkhauser, Boston.

DEMBO, A. and ZEITOUNI, O. (1993). *Large Deviations Techniques and Applications.* Jones and Bartlett, Boston.

DEUSCHEL, J.D. and STROOCK, D.W. (1989). *Large Deviations*, Academic Press, New York.

DINWOODIE, I.H. and ZABELL, S.L. (1992). Large deviations for exchangeable random vectors. *Ann. Probab.* 20 1147-1166.

DUGUNDJI, J. (1970). *Topology*, Allyn and Bacon, Inc., Boston.

EFRON, B. (1979). Bootstrap methods: Another look at the Jackknife. *Ann. Statist.* 7 1-26.

ELLIS, R.S. (1984). Large deviations for a general class of random vectors. *Ann. Probab.* 12 1-12.

ELLIS, R.S. (1985). *Entropy, Large Deviations and Statistical Mechanics.* Springer-Verlag, New York.

Reprinted from

STATISTICS & PROBABILITY LETTERS

Statistics & Probability Letters 34 (1997) 245–250

Bahadur slope of the t-statistic for a contaminated normal

N. Rao Chaganty[a,*], Jayaram Sethuraman[b,1]

[a] *Department of Mathematics & Statistics, Old Dominion University, Norfolk, VA 23529, USA*
[b] *Department of Statistics, Florida State University, Tallahassee, FL 32306, USA*

Received March 1996; revised August 1996

ELSEVIER

ELSEVIER

Statistics & Probability Letters 34 (1997) 245–250

STATISTICS & PROBABILITY LETTERS

Bahadur slope of the t-statistic for a contaminated normal

N. Rao Chaganty[a,*], Jayaram Sethuraman[b,1]

[a] *Department of Mathematics & Statistics, Old Dominion University, Norfolk, VA 23529, USA*
[b] *Department of Statistics, Florida State University, Tallahassee, FL 32306, USA*

Received March 1996; revised August 1996

Abstract

In this paper we derive the exact Bahadur slope of the t-statistic based on a random sample from a contaminated normal distribution, using some results in large deviation theory. We also present a table of exact Bahadur slopes at various alternatives at several levels of contamination.

AMS classification: primary 62F03; 62F05; 62G35; secondary 60F10

Keywords: Bahadur slope; Large deviations; Robustness; Tukey model

1. Introduction

To study robustness of standard tests of location in a normal model, one generally studies their properties under the Tukey model (see Tukey, 1960) of contaminated normal alternatives, namely, the probability distributions $P_{(\varepsilon,\theta,\sigma)}$ with probability density function (pdf)

$$f_{(\varepsilon,\theta,\sigma)}(x) = (1-\varepsilon)\phi(x;\theta,1) + \varepsilon\phi(x;\theta,\sigma) \tag{1}$$

for $0 < \varepsilon < 1$, where $\phi(x;\theta,\sigma)$ is the pdf of a normal distribution with mean θ and variance σ^2.

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from $f_{(\varepsilon,\theta,\sigma)}(x)$ and that we wish to test the null hypothesis $\theta = 0$ versus $\theta > 0$, using the t-statistic $T_n = \sqrt{n}\,\bar{X}_n/S_n$, where $\bar{X}_n = (1/n)\sum_{i=1}^n X_i$ and $S_n^2 = (1/n)\sum_{i=1}^n (X_i - \bar{X}_n)^2$. The robustness of this t-test as measured by Pitman efficiency has been studied in the famous Princeton study by Andrews et al. (1972). In this paper we derive the large deviation rate function of T_n under $P_{(\varepsilon,0,\sigma)}$ which allows us to obtain the exact Bahadur slope of the t-test under a general alternative $P_{(\varepsilon,\theta,\sigma)}$, $\theta > 0$. Following the practice of other authors, we set σ equal to 3, and give the exact Bahadur slopes for various values of ε and θ in Table 1. This table gives an indication of the region of robustness of the t-test as measured by the exact Bahadur slope. The robustness of the t-test, in the sense of Bahadur efficiency, is

* Corresponding author. Partially supported by the US Army research office grant no. DAAH04-96-1-0070.
[1] Partially supported by the US Army research office grant no. DAAH04-93-G-0201.

PII S0167-7152(96)00187-3

gleaned by comparing the slope at the contaminated distribution $P_{(\varepsilon,\theta,3)}$ with the slope at the uncontaminated distribution $P_{(0,\theta,3)}$. As expected, Table 1 shows that there is adequate robustness in a region of small values of ε. Furthermore, for a fixed θ the slope is a decreasing function of ε and for a fixed ε the slope is an increasing function of θ.

The exact distribution of T_n^2 under $P_{(\varepsilon,\theta,\sigma)}$ has been derived in Lee and Gurland (1977). We will derive the large deviation rate function of T_n under $P_{(\varepsilon,0,\sigma)}$ and the exact Bahadur slope under the alternative $P_{(\varepsilon,\theta,\sigma)}$ in Section 2.

2. Large deviation rates and Bahadur slopes

We refer to the excellent monograph of Varadhan (1984) for an introduction to the theory of large deviations and to the monograph of Bahadur (1971) for the concept of Bahadur slopes and efficiencies. One needs a strong law under the alternative and a large deviation result under the null hypothesis to obtain the exact Bahadur slope. It is easy to see from the usual strong law of large numbers that

$$\frac{T_n}{\sqrt{n}} \to m(\varepsilon,\theta,\sigma) = \frac{\theta}{\sqrt{(1-\varepsilon)+\varepsilon\sigma^2}}, \tag{2}$$

with probability one under $P_{(\varepsilon,\theta,\sigma)}$. We need to obtain a result of the form

$$\frac{1}{n}\log P_{(\varepsilon,0,\sigma)}\left(\frac{T_n}{\sqrt{n}} \geqslant m\right) \to -\gamma(m), \tag{3}$$

where $\gamma(m)$ is continuous in m, which is usually referred to as the large deviation rate function of T_n. It then follows that the exact Bahadur slope of T_n equals

$$c(\varepsilon,\theta,\sigma) = 2\gamma(m(\varepsilon,\theta,\sigma)). \tag{4}$$

We now proceed with the derivation of $\gamma(m)$. Note that the event $\{T_n^2/n \geqslant m^2\}$ is equal to the event $\{W_n \geqslant 0\}$ where W_n is the quadratic form $W_n = X'AX/n$ with $A = J - naI$, $a = m^2/(1+m^2)$, I is the identity matrix and J is a matrix of ones. Since the distribution of T_n is symmetric under $P_{(\varepsilon,0,\sigma)}$, we have

$$P\left(\frac{T_n}{\sqrt{n}} \geqslant m\right) = \frac{1}{2}P(W_n \geqslant 0). \tag{5}$$

(From here onwards, P without a suffix corresponds to the probability under $P_{(\varepsilon,0,\sigma)}$.) The logarithm of the probability in (5) can be approximated (see (19) and (20) below) by using the moment generating function (mgf) of W_n which is given by

$$M_n(t) = E[\exp(tW_n)] = \sum_{k=0}^{n}\binom{n}{k}(1-\varepsilon)^k\varepsilon^{(n-k)}\left|I - \frac{2t}{n}A_kA\right|^{-1/2}, \tag{6}$$

where $A_k = \operatorname{diag}(\overbrace{1,\ldots,1}^{k},\overbrace{\sigma^2,\ldots,\sigma^2}^{n-k})$. Let $p = k/n$ and $q = 1-p$. Using a matrix determinant formula (see the appendix), we can show that

$$M_{nk}(t) = \left|I - \frac{2t}{n}A_kA\right|^{-1/2} = (f_1(t))^{-np/2}(f_2(t))^{-nq/2}\left(\frac{pf_2(t)f_3(t)+qf_1(t)f_4(t)}{f_1(t)f_2(t)}\right)^{-1/2}, \tag{7}$$

where $f_1(t) = 1 + 2at$, $f_2(t) = 1 + 2at\sigma^2$, $f_3(t) = 1 - 2t(1-a)$ and $f_4(t) = 1 - 2t\sigma^2(1-a)$. Thus, the mgf of W_n is given by

$$M_n(t) = \sum_{k=0}^{n} \binom{n}{k} (1-\varepsilon)^k \varepsilon^{(n-k)} M_{nk}(t) \quad \text{for } t_*(p) < t < t^*(p), \tag{8}$$

where $t_*(p)$, $t^*(p)$ are the roots of the quadratic equation $pf_2(t)f_3(t) + qf_1(t)f_4(t) = 0$.

From the above formula for the mgf $M_n(t)$, we can conclude that the distribution of W_n is a mixture distribution. More precisely, let K be a binomial random variable with parameters n and $(1-\varepsilon)$. Given $K = k$, let U_{nk} be a random variable with mgf given by M_{nk}. From (8) we can see that W_n is equal in distribution to U_{nK}. This observation coupled with a theorem of Varadhan, see Theorem 2.2 in Chaganty (1993), is useful to derive the large deviation rate function for the random variable W_n. Theorem 1 below shows that the conditions in Varadhan's theorem are indeed satisfied in our problem.

Theorem 1. *Let K be a binomial random variable with parameters n and $(1-\varepsilon)$. Given $K = k_n = np_n$, let U_{nk_n} be a random variable with mgf, $M_{nk_n}(t)$, defined in (7). If $p_n \to p$ then*

$$F_n(p_n) = \frac{1}{n} \log P(U_{nk_n} \geqslant 0) \to F(p) \quad \text{as } n \to \infty, \tag{9}$$

where $F(p) = -\frac{1}{2}[p \log f_1(t^*(p)) + q \log f_2(t^*(p))]$, $q = 1 - p$.

Proof. *Upper bound*: By Chebyshev's inequality it follows that

$$\begin{aligned} \limsup_n \frac{1}{n} \log P(U_{nk_n} \geqslant 0) &\leqslant \lim_n \frac{1}{n} \log M_{nk_n}(t) \\ &= -\frac{1}{2}[p \log f_1(t) + q \log f_2(t)] \end{aligned} \tag{10}$$

for any $0 < t < t^*(p)$. Hence,

$$\begin{aligned} \limsup_n F_n(p_n) &= \limsup_n \frac{1}{n} \log P(U_{nk_n} \geqslant 0) \\ &\leqslant \inf_{0<t<t^*(p)} -\frac{1}{2}[p \log f_1(t) + q \log f_2(t)] \\ &= F(p). \end{aligned} \tag{11}$$

Lower bound: Let G_{nk_n} denote the distribution function of U_{nk_n}. Let us introduce another random variable V_n with the conjugate distribution function given by

$$\mathrm{d}H_{nt_n}(x) = \frac{\exp(xt_n)}{M_{nk_n}(t_n)} \mathrm{d}G_{nk_n}(x) \tag{12}$$

where $t_n = t^*(p)(1 - (1/n))$. Now for any $\delta > 0$ we have

$$\begin{aligned} P(U_{nk_n} \geqslant 0) &= \int_0^\infty \mathrm{d}G_{nk_n}(x) = M_{nk_n}(t_n) \int_0^\infty \exp(-xt_n)\, \mathrm{d}H_{nt_n}(x) \\ &\geqslant M_{nk_n}(t_n) \int_0^{n\delta} \exp(-xt_n)\, \mathrm{d}H_{nt_n}(x) \\ &\geqslant M_{nk_n}(t_n) \exp(-n\delta t_n) P(0 \leqslant V_n \leqslant n\delta). \end{aligned} \tag{13}$$

Therefore,

$$\frac{1}{n}\log P(U_{nk_n}\geqslant 0)\geqslant \frac{1}{n}\log M_{nk_n}(t_n)-\delta t_n+\frac{1}{n}\log P(0\leqslant V_n\leqslant n\delta). \tag{14}$$

Since $p_n\to p$ as $n\to\infty$ it follows from (7)

$$\frac{1}{n}\log M_{nk_n}(t_n)\to -\frac{1}{2}\left[p\log f_1(t^*(p))+q\log f_2(t^*(p))\right]=F(p). \tag{15}$$

We will now show that the limiting distribution of V_n/n is a translated gamma distribution. To find the limiting distribution, we first note that the mgf of V_n/n is given by $M_n(s)=M_{nk_n}(s_n)/M_{nk_n}(t_n)$, where $s_n=t_n+s/n$. It is easy to check that

$$M_n(s)\to M(s)=\exp(-sc)\left(\frac{t^*(p)}{t^*(p)-s}\right)^{1/2}\quad \text{as } n\to\infty, \tag{16}$$

for $s<t^*(p)$, where $c=[ap/(1+2at^*(p))+aq\sigma^2/(1+2at^*(p)\sigma^2)]$. Thus, V_n/n converges in distribution to $V-c$, where V is a Gamma random variable with shape parameter $1/2$ and scale parameter $1/t^*(p)$. Therefore,

$$P(0\leqslant V_n/n\leqslant\delta)\to P(c\leqslant V\leqslant c+\delta)>0\quad \text{as } n\to\infty. \tag{17}$$

From (14), (15) and (17) we get

$$\liminf_n F_n(p_n)=\liminf_n \frac{1}{n}\log P(W_{nk_n}\geqslant 0)\geqslant F(p)-\delta t^*(p).$$

Since δ is arbitrary we get $\liminf_n F_n(p_n)\geqslant F(p)$. This completes the proof of the theorem. □

We are now in a position to derive the large deviation rate function $\gamma(m)$ of T_n. From Theorem 1 we have,

$$F_n(p_n)=\frac{1}{n}\log P(W_n\geqslant 0\,|\,K=np_n)\to F(p) \tag{18}$$

whenever $p_n\to p$. Note that

$$\frac{1}{n}\log P(W_n\geqslant 0)=\frac{1}{n}\log\int \exp(nF_n(p))\,d\mu_n(p), \tag{19}$$

where μ_n is the distribution of K/n. Since the distribution of K is binomial, it is known that the sequence of probability measures $\{\mu_n\}$ obeys the large deviation principle (see Varadhan, 1984 for the definition) with rate function

$$h(p)=p\log(p/(1-\varepsilon))+q\log(q/\varepsilon).$$

Using the theorem of Varadhan, see Theorem 2.2 in Chaganty (1993), and (18) and (19) it follows that

$$\frac{1}{n}\log P(W_n\geqslant 0)\to \sup_{0<p<1}(F(p)-h(p)). \tag{20}$$

From (5) and (20) we get

$$\frac{1}{n}\log P\left(\frac{T_n}{\sqrt{n}}\geqslant m\right)\to -\gamma(m),$$

where $\gamma(m)=\inf_{0<p<1}[-F(p)+h(p)]$.

Table 1
Slope of the t-statistic $c(\varepsilon, \theta, \sigma)$, for the contaminated normal model, when $\sigma = 3$

ε	$\theta = 0.25$	$\theta = 0.50$	$\theta = 1.0$	$\theta = 1.5$	$\theta = 2.0$	$\theta = 2.5$	$\theta = 3.0$
0.00	0.06062	0.22314	0.69315	1.17865	1.60944	1.98100	2.30259
0.05	0.04487	0.17380	0.56738	0.99565	1.39154	1.74207	2.05046
0.10	0.03509	0.14056	0.48860	0.87952	1.24944	1.58306	1.88034
0.15	0.02865	0.11598	0.42937	0.79694	1.14852	1.46908	1.75733
0.25	0.02090	0.08422	0.33264	0.67161	1.00633	1.31239	1.58918

The rate function $\gamma(m)$ can easily be computed numerically using Newton–Raphson method. In Table 1 we present the exact Bahadur slope, $c(\varepsilon, \theta, \sigma) = 2\gamma(m(\varepsilon, \theta, \sigma))$, for different values of ε and θ when $\sigma = 3$. Note that a large value of $c(\varepsilon, \theta, \sigma)$ indicates that the test statistic T_n requires smaller sample size to detect that particular alternative. The Bahadur efficiency of the t-test with respect to the competing nonparametric Wilcoxon test in the Tukey model has recently been obtained in Chaganty and Sethuraman (1996).

Remark 1. It is possible to derive, in a similar manner, the exact Bahadur slope of the t-statistic, for a random sample of n observations with common pdf given by $f(x) = \sum_{i=1}^{L} \pi_i \phi(x; \theta, \sigma_i)$, $\sum_{i=1}^{L} \pi_i = 1$, and $\pi_i > 0$ for all $L \geqslant 1$. In this case the multinomial distribution plays the role of the binomial distribution in the derivation of the slope. More generally, using the results of Chaganty (1993), we can also establish the large deviation principle for the t-statistic for this model.

Appendix

In (7) we have used the following determinant formula. Let

$$S = \overbrace{\begin{bmatrix} bI + cJ & cJ \\ eJ & dI + eJ \end{bmatrix}}^{k \qquad (n-k)},$$

where b, c, d and e are constants, and as before, I is the identity matrix and J is the matrix of ones. Then we can verify that

$$|S| = b^k d^{(n-k)} \left(1 + \frac{kc}{b} + \frac{(n-k)e}{d}\right). \tag{A.1}$$

To obtain the simplification in Eq. (7), we use the above formula (A.1) with the substitutions $b = f_1(t)$, $d = f_2(t)$, $c = -2t/n$ and $e = -2t\sigma^2/n$.

References

Andrews, D.F., P.J. Bickel, F.R. Hampel, P.J. Huber, W.H. Rogers and J.W. Tukey (1972), *Robust Estimates of Location: Survey and Advances* (Princeton Univ. Press, Princeton, NJ).

Bahadur, R.R. (1971), *Some Limit Theorems in Statistics*, SIAM, CBMS/NSF Regional Conf. in Applied Math., Vol. 4 (SIAM, Philadelphia).

Chaganty N.R. (1993), Large deviations for joint distributions and statistical applications, Tech. Report #TR93-2, Department of Mathematics & Statistics, Old Dominion University, Norfolk.

Chaganty N.R. and J. Sethuraman (1996), The large deviation principle for common statistical tests against a contaminated normal, Tech. Report #M909, Department of Statistics, Florida State University, Tallahassee, FL.

Lee, A.F.S. and J. Gurland (1977), One sample t-test when sampling from a mixture of normal distributions, *Ann. Statist.* **5**, 803–807.
Tukey, J.W. (1960), A survey of sampling from contaminated distributions, in: I. Olkin, ed., *Contributions to Probability and Statistics* (Stanford Univ. Press, Stanford, CA).
Varadhan, S.R.S. (1984), *Large Deviations and Applications*, SIAM, CBMS/NSF Regional Conf. in Applied Math., Vol. 46 (SIAM, Philadelphia).

16

The Large Deviation Principle for Common Statistical Tests Against a Contaminated Normal

N. Rao Chaganty and J. Sethuraman

Old Dominion University, Norfolk, VA
Florida State University, Tallahassee, FL

Abstract: We examine the performance of the standard tests—the mean test, the t-test, the Wilcoxon test and the sign test—for testing that the measure of central tendency of a distribution is zero. We do this by comparing the Bahadur slopes in a contaminated normal model. We first establish the large deviation principle (LDP) and then calculate the Bahadur slopes for the standard test statistics when the observations come from a contaminated normal distribution. An examination of tables of Bahadur efficiencies reveals that the Wilcoxon test outperforms other tests in a neighborhood of the null hypothesis, even in the presence of moderate contamination, but not uniformly over the whole alternative hypothesis.

Keywords and phrases: Bahadur slope, large deviations, Pitman efficiency, robustness, Tukey model. Wilcoxon test

16.1 Introduction

One of the most common testing problems encountered in statistics is testing

$$H_0 : \theta = 0 \qquad vs. \qquad H_1 : \theta > 0$$

where θ is a measure of central tendency. For simplicity, one makes the assumption that the sample forms an i.i.d. sample from a normal distribution with unknown variance. In this case, the t-test is known to be the uniformly most powerful unbiased test. Other tests that have been proposed include the mean test, the sign test, and the Wilcoxon test. Examining the robustness of these tests against departures from this model has been the subject of a large number of papers; see Staudte (1980) and the books by Andrews *et al.* (1972),

Huber (1981) and Tiku, Tan and Balakrishnan (1986), and more recently by DasGupta (1994). It has been the standard practice to examine the robustness of these tests in the famous Tukey model [see Tukey (1960)], which models a certain form of departure from normality. Under the Tukey model, the sample consists of i.i.d. observations from the density

$$f_{\theta,\epsilon,\sigma}(x) = (1-\epsilon)\,\phi(x;\theta,1) + \epsilon\,\phi(x;\theta,\sigma). \tag{16.1}$$

Here $\phi(x;\theta,\sigma)$ denotes the probability density function of a normal random variable with mean θ and standard deviation σ, and $\epsilon \in (0,1)$ represents the level of contamination.

Two measures which are commonly used to compare the large sample performance of tests are Pitman efficiency and Bahadur efficiency. In Andrews *et al.* (1972), Huber (1981) and Lehmann (1983, Chapter 5), robustness was measured by Pitman efficiency, which is obtainable by comparing asymptotic efficacies of tests. In this paper, we measure the robustness of these tests by Bahadur efficiency, which is obtainable by comparing Bahadur slopes. We present some tables showing the Bahadur efficiencies of the Wilcoxon test relative to other three tests. From an examination of these tables, it appears that the Wilcoxon test is the best performer in a neighborhood of the null hypothesis, even under the presence of moderate contamination, but is not the best performer uniformly over the whole region of the alternative hypothesis.

The concept of Bahadur slope can be briefly described as follows. Let $X_1, \ldots, X_n$ be i.i.d., whose distribution depends on a parameter λ taking values in a set Λ. The parameter λ can be a vector like $(\theta, \epsilon, \sigma)$ as occurs in our problem. Consider the problem of testing the hypothesis that λ lies in a subset Λ_0 of Λ. For each n, let T_n be a real valued function of the sample $\{X_1, X_2, \ldots, X_n\}$, such that large values of T_n are significant for testing the null hypothesis. For any λ and t, let

$$F_n(t, \lambda) = P_\lambda(T_n < t) \tag{16.2}$$

and

$$G_n(t) = \inf\{F_n(t, \lambda) : \lambda \in \Lambda_0\}. \tag{16.3}$$

If Λ_0 were a singleton, then $F_n(t, \lambda)$ and $G_n(t)$ are equal; otherwise, the significance probability of a test based on T_n is obtained from $G_n(t)$. In fact, the level attained by T_n is

$$L_n(T_n) = 1 - G_n(T_n). \tag{16.4}$$

The rate at which L_n tends to zero when a non-null λ obtains is a measure of the discriminating power of the sequence of test statistics $\{T_n\}$ in discriminating

that λ; see Bahadur (1960, 1967, 1971). The sequence of test statistics $\{T_n\}$ is said to have exact slope $c(\lambda)$ when λ obtains if

$$\lim_{n\to\infty} \frac{1}{n} \log L_n(T_n) = -\frac{1}{2} c(\lambda) \text{ a.s. } [P_\lambda]. \tag{16.5}$$

It is only the values of $c(\lambda)$ for $\lambda \in \Lambda \setminus \Lambda_0$ that are of interest, with larger values indicating that the alternative hypothesis λ is discriminated better. In general, it is a nontrivial matter to determine the exact slope of a given sequence $\{T_n\}$. A convenient and frequently used method to obtain Bahadur slopes is due to Bahadur (1967) and can be stated in the form of the following theorem.

Theorem 16.1.1 *Suppose that for each* $\lambda \in \Lambda \setminus \Lambda_0$,

$$\lim_{n\to\infty} T_n = b(\lambda) \quad \text{a.s. } [P_\lambda] \tag{16.6}$$

where $-\infty < b(\lambda) < \infty$. *Suppose that for* $\lambda \in \Lambda_0$

$$\lim_{n\to\infty} \frac{1}{n} \log L_n(s) = -I(s) \quad \text{a.s. } [P_\lambda] \tag{16.7}$$

for each s *in an open interval which includes* $b(\lambda)$, *and* $I(s)$ *is a positive continuous function on that interval. Then, the exact slope of* $\{T_n\}$ *exists for each* $\lambda \in \Lambda \setminus \Lambda_0$ *and equals* $c(\lambda) = 2I(b(\lambda))$.

In practice, verification of condition (16.6) is easy and usually follows from a strong law of large numbers. However, a large deviation theorem is needed to establish (16.7) and is usually the difficult part. For sums of i.i.d. random variables, one can use Cramér's or Chernoff's theorem [see Theorem 3.1 of Bahadur (1971), for instance] to establish (16.7). For statistics with completely general structure, it is more convenient to use the main theorem of Ellis (1984) to establish (16.7). We will use both these methods in this paper.

We will now briefly give the definition of the large deviation principle and state the main theorem of Ellis (1984).

A function $I(s) : \mathcal{R}^k \to [0, \infty]$ is said to be a *rate function* if it is lower semi-continuous. For any subset A, we write $I(A) = \inf\{I(s) : s \in A\}$. Let $\{\mu_n\}$ be a sequence of probability measures on $(\mathcal{R}^k, \mathcal{B})$. We say that $\{\mu_n\}$ obeys the *large deviation principle* (LDP) with rate function $I(s)$ [see Varadhan (1984)] if the following conditions are satisfied:

$$\limsup_n \frac{1}{n} \log \mu_n(C) \leq -I(C) \tag{16.8}$$

$$\liminf_n \frac{1}{n} \log \mu_n(G) \geq -I(G) \tag{16.9}$$

for all closed sets C and for all open sets G, respectively, of $\mathcal{R}^k$. The rate function $I(s)$ is said to be a *proper rate function* if it further satisfies the condition that the level set $\{s : I(s) \leq L\}$ is a compact subset of $\mathcal{R}^k$, for each $L \geq 0$.

Let $\{T_n\}$ be a sequence of $\mathcal{R}^k$ valued random variables with distribution given by the sequence of probability measures $\{\mu_n\}$. Define

$$c_n(t) = \frac{1}{n} \log E[\exp(< t, nT_n >)]. \tag{16.10}$$

Suppose that $\lim_n c_n(t)$ exists and is equal to $c(t)$, for all $t \in \mathcal{R}^k$, where we allow both $c_n(t)$ and $c(t)$ to take the value $+\infty$. Let $\mathcal{D}(c) = \{t \in \mathcal{R}^k : c(t) < \infty\}$. The function $c(t) : \mathcal{R}^k \to \mathcal{R}$ is said to be closed if $\{t : c(t) \leq \alpha\}$ is closed for each real α. This is equivalent to $c(t)$ being lower semi-continuous. If $c(t)$ is differentiable on the interior of $\mathcal{D}(c)$, then we call $c(t)$ steep if $\| \text{ grad}(t_n)) \| \to \infty$ for any sequence $\{t_n\} \subset \text{ int } (\mathcal{D}(c))$ which tends to a boundary point of $\mathcal{D}(c)$. Let

$$I(s) = \sup_{t \in \mathcal{R}^k} [< t, s > - c(t)], \tag{16.11}$$

for $s \in \mathcal{R}^k$ be the Legendre-Fenchel transform of $c(t)$. The main theorem of Ellis (1984) can then be stated as follows.

Theorem 16.1.2 (Ellis): *If $\mathcal{D}(c)$ has a nonempty interior containing the point t=0 and $c(t)$ is a closed convex function of $\mathcal{R}^k$, then the function $I(s)$ defined in (16.11) is a proper rate function on $\mathcal{R}^k$ and the sequence of probability measures $\{\mu_n\}$ satisfies the upper bound (16.8) for all closed sets C of $\mathcal{R}^k$ with proper rate function $I(s)$. Furthermore, if $c(t)$ is differentiable on all of interior of $\mathcal{D}(c)$ and is steep, then the sequence of probability measures $\{\mu_n\}$ satisfies the lower bound (16.9) for all open sets G of $\mathcal{R}^k$ with proper rate function $I(s)$.*

16.2 LDP for Common Statistical Tests

Let $X_1, X_2, \ldots, X_n$ be a random sample from the distribution (16.1). In this section, we will first establish the LDP for the commonly used test statistics for testing the hypothesis $H_0 : \theta = 0$ *vs.* $H_1 : \theta > 0$. The LDP results for the Wilcoxon and t-statistics are new. We do this even though the full force of the LDP is not required to calculate Bahadur slopes.

We will consider four test statistics—the mean test, the t-test, the Wilcoxon test, and the sign test—the last two of which are nonparametric tests.

Mean test. The test statistic (under the assumption that the population variance is known) for the mean test is $T_{1n} = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Under the null hypothesis $H_0 : \theta = 0$, we have

$$\begin{aligned} c_{1n}(t) &= \frac{1}{n}\log E[\exp(n\,t\,T_{1n})] \\ &= \frac{t^2}{2} + \log\left[(1-\epsilon) + \epsilon\,\exp\left(\frac{t^2(\sigma^2-1)}{2}\right)\right] \\ &= c_1(t) \quad (\text{say}), \quad -\infty < t < \infty, \end{aligned} \tag{16.12}$$

is independent of n. It is easy to verify that the function $c_1(t)$ is a closed convex function on the real line, and satisfies the hypothesis of Theorem 16.1.2. Therefore, T_{1n} obeys the LDP with proper rate function

$$\begin{aligned} I_1(s) &= \sup_{-\infty<t<\infty}[s\,t - c_1(t)] \\ &= s\,t_s - c_1(t_s) \end{aligned} \tag{16.13}$$

where t_s satisfies the equation $s = c_1'(t_s)$, which simplifies to

$$e^{t_s^2/2}\,(1-\epsilon)(s-t_s) + \epsilon\,e^{\sigma^2 t_s^2/2}\,(s - t_s\sigma^2) = 0. \tag{16.14}$$

The above equation (16.14) can be solved numerically using the Newton-Raphson method.

Sign test. Let $Y_i = 1$ if $X_i > 0$ and $Y_i = 0$ if $X_i \le 0$. The nonparametric sign test is based on the statistic $T_{2n} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Note that the random variables Y_i's are i.i.d. Bernoulli with mean $1/2$ under the null hypothesis $H_0 : \theta = 0$. It is well known, and can also be alternatively derived from Theorem 16.1.2, that T_{2n} obeys the LDP with proper rate function given by

$$I_2(s) = \begin{cases} \log(2) + s\log(s) + (1-s)\log(1-s). & \text{if } 0 \le s \le 1 \\ \infty & \text{otherwise.} \end{cases} \tag{16.15}$$

Wilcoxon test. Arrange $|X_1|, \ldots, |X_n|$ in increasing order and assign ranks. Let U_i be the sign of X_j where $|X_j|$ has rank i. The Wilcoxon statistic is equivalent to

$$T_{3n} = \frac{1}{n(n+1)}\sum_{i=1}^{n} i\,U_i\,. \tag{16.16}$$

The following theorem generalizes a result of Klotz (1965) and gives the LDP for the Wilcoxon statistic.

Theorem 16.2.1 *Let E_{ni} denote the expected value of the ith smallest order statistic from a sample of n observations with distribution function G on $(0, \infty)$ satisfying $\int_0^\infty x\, dG(x) < \infty$. Let U_i be independent such that $P(U_i = \pm) = 1/2$ for $i = 1, 2, \ldots, n$ and let $S_n = \frac{1}{n}\sum_{i=1}^{n} E_{ni}\, U_i$. Then, $\{S_n\}$ obeys the LDP with proper rate function given by*

$$I(s) = \sup_t \left[st - \int_0^\infty \log(\cosh(xt))\, dG(x) \right]. \tag{16.17}$$

PROOF. From Theorem 1 of Hoeffding (1953), we have for each $t \in \mathcal{R}$

$$\begin{aligned} c_{3n}(t) &= \frac{1}{n} \log E[\exp(n\, t\, S_n)] \\ &= \frac{1}{n} \sum_{i=1}^{n} \log[\cosh(t\, E_{ni})] \\ &\to \int_0^\infty \log(\cosh(t\, x))\, dG(x) = c_3(t) \quad \text{(say)} \end{aligned} \tag{16.18}$$

as $n \to \infty$. It is easy to check that the function $c_3(t)$ satisfies the conditions of Theorem 16.1.2. Theorem 16.2.1 now follows from Theorem 16.1.2. ■

Note that under the null hypothesis $\theta = 0$, the random variables $U_1, \ldots, U_n$ in (16.16) are i.i.d. symmetric Bernoulli. Also in (16.16), $E_{ni} = \frac{i}{n+1}$ is the expected value of the ith smallest order statistic from a random sample of n observations distributed uniformly on (0, 1). Therefore, conditions of Theorem 16.2.1 apply with G as the *uniform cdf* on (0, 1). Let

$$c_3(t) = \int_0^1 \log[\cosh(t\, x)]\, dx, \qquad -\infty < t < \infty. \tag{16.19}$$

Using Theorem 16.2.1, we can conclude that T_{3n} obeys the LDP with proper rate function

$$\begin{aligned} I_3(s) &= \sup_t [s\, t - c_3(t)] \\ &= s\, t_s - c_3(t_s) \end{aligned} \tag{16.20}$$

where t_s is the solution of the equation

$$\begin{aligned} s &= c_3'(t) \\ &= \int_0^1 x \tanh(t\, x)\, dx \\ &= \frac{1}{2} - \frac{\pi^2}{24\, t^2} + \frac{\log(1 + \exp(-2t))}{t} + \frac{1}{2t^2} \sum_{k=1}^{\infty} (-1)^{k+1} \frac{\exp(-2tk)}{k^2}. \end{aligned} \tag{16.21}$$

Equation (16.21) can be solved numerically using the Newton-Raphson method. By substituting an alternate expression for $c_3(t)$ obtainable from (16.19) using integration by parts, we can rewrite (16.20) as

$$I_3(s) = 2s\,t_s - \log(\cosh(t_s)) \tag{16.22}$$

where t_s is the solution of the equation (16.21).

t-test. Let $\overline{X}_n$ and $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$ be the mean and variance of the sample. The t-statistic is simply defined as $T_{4n} = \overline{X}_n / S_n$. The LDP for the t-statistic does not follow from Theorem 16.1.2. However, we can establish the LDP for the t-statistic using a recent large deviation theorem of Chaganty (1997). Let K_n be distributed as Binomial with parameters n and $(1-\epsilon)$. Let $Z_1, Z_2, \ldots$ be i.i.d. $N(0, 1)$ and $Y_1, Y_2, \ldots$ be i.i.d. $N(0, \sigma)$, independent of K_n. Let $\overline{Z}_n$ and $\overline{Y}_n$ be the sample means and let S_{zn}^2 and S_{yn}^2 be the sample variances of a sample of n observations from Z and Y, respectively. Note that T_{4n} is equal in distribution to the statistic

$$\frac{P_n\,\overline{Z}_{nP_n} + (1-P_n)\,\overline{Y}_{n(1-P_n)}}{\sqrt{P_n\,S^2_{z\,nP_n} + (1-P_n)\,S^2_{y\,n(1-P_n)} + P_n(1-P_n)(\overline{Z}_{nP_n} - \overline{Y}_{n(1-P_n)})^2}}$$

where $P_n = K_n/n$. It is well known that $\overline{Z}_n$ obeys the LDP with proper rate function $h_1(z) = z^2/2$ and $\overline{Y}_n$ obeys the LDP with proper rate function $h_2(y) = y^2/2\sigma^2$. Also, S_{zn}^2 and S_{yn}^2 obey the LDP with proper rate functions $h_3(u) = [u - 1 - \log(u)]/2$ and $h_4(v) = [v/\sigma^2 - 1 - \log(v/\sigma^2)]/2$, respectively. Conditional on $P_n = p$, the variables $\{\overline{Z}_{nP_n}, \overline{Y}_{n(1-P_n)}, S^2_{z\,nP_n}, S^2_{y\,n(1-P_n)}\}$ are all independent. Using Corollary 2.9 in Lynch and Sethuraman (1987) and Example 3.11 in Chaganty (1997), we can see that this conditional joint distribution obeys the LDP continuity condition in p with proper rate function given by $h_1(z) + h_2(y) + h_3(u) + h_4(v)$. See Chaganty (1997) for the definition of the LDP continuity condition and the contraction principle in that connection. It follows from that contraction principle that the conditional distribution of T_{4n} given $P_n = p$ also satisfies the LDP continuity condition in p with proper rate function given by

$$J(p, s) = \inf_{(z,y,u,v)\,:\,s = \frac{pz+(1-p)y}{\sqrt{pu+qv+p(1-p)(z-y)^2}}} [h_1(z) + h_2(y) + h_3(u) + h_4(v)]. \tag{16.23}$$

From the LDP for binomial distributions, P_n obeys the LDP with proper rate function given by $h_5(p) = p\log(p/(1-\epsilon)) + (1-p)\log((1-p)/\epsilon)$. It then follows from Theorem 2.3 of Chaganty (1997) that T_{4n} obeys the LDP with proper rate function

$$I_4(s) = \inf_{0<p<1} [J(p, s) + h_5(p)] \tag{16.24}$$

where $J(p, s)$ is given by (16.23). The above expression (16.24) for $I_4(s)$ is not convenient for computational purposes. However, using a different approach, Chaganty and Sethuraman (1997) have derived the following equivalent form for the rate function

$$I_4(s) = \inf_{0<p<1} \frac{1}{2}\left[p\log(1+2at^*) + (1-p)\log(1+2a\sigma^2 t^*) + 2h_5(p)\right], \tag{16.25}$$

where

$$a = s^2/(1+s^2)$$

$$t^* = \frac{\sigma^2(p+a-1) + (a-p) + \sqrt{(\sigma^2(p+a-1)+a-p)^2 + 4\sigma^2 a(1-a)}}{4\sigma^2 a(1-a)}.$$

We use the expression in (16.25) in our calculations.

16.3 Bahadur Slopes and Efficiencies

We now derive the Bahadur slopes of the common test statistics for testing $H_0 : \theta = 0$ *vs.* $H_1 : \theta > 0$ in the Tukey model, using the results of Section 16.2. These slopes will depend on the alternative hypothesis, i.e., on the vector $\lambda = (\theta, \epsilon, \sigma)$ with $\theta > 0$.

1. The Bahadur slope of the mean test is

$$c_m(\lambda) = 2[\theta t_\lambda - c_1(t_\lambda)] \tag{16.26}$$

where $c_1(t)$ is defined in (16.12) and t_λ satisfies the equation

$$e^{t_\lambda^2/2}(1-\epsilon)(\theta - t_\lambda) + \epsilon e^{\sigma^2 t_\lambda^2/2}(\theta - t_\lambda \sigma^2) = 0. \tag{16.27}$$

2. The Bahadur slope of the sign test is

$$c_s(\lambda) = 2[\log 2 + p_\lambda \log(p_\lambda) + q_\lambda \log(q_\lambda)] \tag{16.28}$$

where $p_\lambda = (1-\epsilon)\,\Phi(\theta) + \epsilon\,\Phi(\theta/\sigma)$, $q_\lambda = 1 - p_\lambda$ and Φ is the cdf of the standard normal distribution.

3. The Bahadur slope of the Wilcoxon test is

$$c_w(\lambda) = 2\,[2\,b(\lambda)\,t_\lambda - \log(\cosh(t_\lambda))] \tag{16.29}$$

where

$$b(\lambda) = \epsilon^2\Phi\left(\frac{\sqrt{2}\,\theta}{\sigma}\right) + 2\epsilon(1-\epsilon)\Phi\left(\frac{2\theta}{\sqrt{1+\sigma^2}}\right) + (1-\epsilon)^2\Phi(\sqrt{2}\theta) - \frac{1}{2} \tag{16.30}$$

and t_λ is the solution of the equation (16.21) with $s = b(\lambda)$.

4. The Bahadur slope of the t-statistic is

$$c_t(\lambda) = 2\, I_4(b(\lambda)) \tag{16.31}$$

where $b(\lambda) = \theta/\sqrt{(1-\epsilon)+\epsilon\,\sigma^2}$ and $I_4(s)$ is given by (16.24).

The Bahadur efficiencies of the mean test, t-test, and the sign test with respect to the Wilcoxon test are defined as the ratio of the slopes and they are given by $e_\lambda(m, w) = c_m(\lambda)/c_w(\lambda)$, $e_\lambda(t, w) = c_t(\lambda)/c_w(\lambda)$ and $e_\lambda(s, w) = c_s(\lambda)/c_w(\lambda)$, respectively. The Pitman efficiencies of these test statistics can be obtained from more general formulas given in Serfling (1980, p. 321); see also Hodges and Lehmann (1956, 1961). In our problem, the Pitman efficiencies of the mean test and the t-test with respect to the Wilcoxon test are equal and the common value is given by

$$\begin{aligned} e_{p\lambda}(m, w) &= e_{p\lambda}(t, w) \\ &= \frac{\pi}{3}[(1-\epsilon)+\epsilon\,\sigma^2]^{-1}\left[(1-\epsilon)^2 + \frac{2\sqrt{2}\epsilon(1-\epsilon)}{\sqrt{1+\sigma^2}} + \frac{\epsilon^2}{\sigma}\right]^{-2} \end{aligned} \tag{16.32}$$

whereas the Pitman efficiency of the sign test with respect to the Wilcoxon test is

$$e_{p\lambda}(s, w) = \frac{2}{3}[(1-\epsilon)+\epsilon/\sigma]^2\left[(1-\epsilon)^2 + \frac{2\sqrt{2}\epsilon(1-\epsilon)}{\sqrt{1+\sigma^2}} + \frac{\epsilon^2}{\sigma}\right]^{-2}. \tag{16.33}$$

Following the convention set in Andrews *et al.* (1972), we have set $\sigma = 3$ and computed the Bahadur efficiencies $e_\lambda(m, w)$, $e_\lambda(t, w)$ and $e_\lambda(s, w)$. These efficiencies will depend on the alternative θ, and the level of contamination ϵ. For simplicity in notation, we shall drop the subscript λ and denote them simply as $e(m, w)$, $e(t, w)$ and $e(s, w)$.

Figures 16.1, 16.2 and 16.3 give the surface of the Bahadur efficiencies $e(m, w)$, $e(t, w)$ and $e(s, w)$ as a function of θ and ϵ. Tables 16.1, 16.2 and 16.3 can be used to view the same information by looking at the performances of the mean test, the t-test and the sign test, simultaneously, with respect to the Wilcoxon test for a fixed level of contamination and varying values of θ. From

a numerical point of view, the corresponding Pitman efficiencies are given by the restriction of these surfaces to the plane $\theta = 0$. The fact that the limiting Bahadur efficiency as $\theta \to 0$ yields the Pitman efficiency has been established in great generality in Wieand (1976), and we conjecture that it is true in this case also. It is clear that the Bahadur efficiencies of the mean test, t-test and the sign test with respect to the Wilcoxon test is less than 1 in a neighborhood of $\theta = 0$ and $\epsilon = 0$, but not on the whole region of alternatives. This leads us to the conclusion that the Wilcoxon test outperforms the remaining tests in a neighborhood of the null hypothesis even under the presence of moderate contamination.

Acknowledgements. Research partially supported by the U.S. Army research office grant numbers DAAH04-96-1-0070 (first author) and DAAH04-93-G-0201 (second author). The United States Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon.

References

1. Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H. and Tukey, J. W. (1972). *Robust Estimates of Location: Survey and Advances*, Princeton: Princeton University Press.

2. Bahadur, R. R. (1960). Stochastic comparison of tests, *Annals of Mathematical Statistics*, **31**, 276–295.

3. Bahadur, R. R. (1967). Rates of convergence of estimates and test statistics, *Annals of Mathematical Statistics*, **38**, 303–324.

4. Bahadur, R. R. (1971). *Some Limit Theorems in Statistics*, CBMS/NSF Regional Conference in Applied Mathematics–4, Philadelphia: SIAM.

5. Chaganty N. R. (1997). Large deviations for joint distributions and statistical applications, *Sankhyā, Series A* (to appear).

6. Chaganty, N. R. and Sethuraman, J. (1997). Bahadur slope of the t-statistic for a contaminated normal, *Statistics & Probability Letters* (to appear).

7. DasGupta, A. (1994). Bounds on asymptotic relative efficiencies of robust estimates of locations for random contaminations, *Journal of Statistical Planning and Inference*, **41**, 73–93.

8. Ellis, R. S. (1984). Large deviations for a general class of random vectors, *Annals of Probability*, **12**, 1–12.

9. Hodges, J. L., Jr. and Lehmann, E. L. (1956). The efficiency of some nonparametric competitors of the t-test, *Annals of Mathematical Statistics*, **27**, 324–335.

10. Hodges, J. L., Jr. and Lehmann, E. L. (1961). Comparison of the normal scores and Wilcoxon tests, *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 307–317.

11. Hoeffding, W. (1953). On the distribution of the expected values of the order statistics, *Annals of Mathematical Statistics*, **24**, 93–100.

12. Huber, P. J. (1981). *Robust Statistics*, New York: John Wiley & Sons.

13. Klotz, J. (1965). Alternative efficiencies for the signed rank tests, *Annals of Mathematical Statistics*, **36**, 1759–1766.

14. Lehmann, E. L. (1983). *Theory of Point Estimation*, New York: John Wiley & Sons.

15. Lynch, J. and Sethuraman, J. (1987). Large deviations for processes with independent increments, *Annals of Probability*, **15**, 610–627.

16. Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*, New York: John Wiley & Sons.

17. Staudte, R. G., Jr. (1980). Robust estimation, *Queen's Papers in Pure and Applied Mathematics*, No. 53, Queen's University, Kingston, Ontario.

18. Tiku, M. L., Tan, W. Y. and Balakrishnan, N. (1986). *Robust Inference*, New York: Marcel Dekker.

19. Tukey, J. W. (1960). A survey of sampling from contaminated distributions, In *Contributions to Probability and Statistics* (Ed., I. Olkin), Stanford, CA: Stanford University Press.

20. Varadhan, S. R. S. (1984). *Large Deviations and Applications*, CBMS/NSF Regional Conference in Applied Mathematics–**46**, Philadelphia: SIAM.

21. Wieand, H. S. (1976). A condition under which the Pitman and Bahadur approaches to efficiency coincide, *Annals of Statistics*, **4**, 1003–1011.

Table 16.1: Bahadur efficiencies of the mean test, t-test and the sign test with respect to the Wilcoxon test when the level of contamination is 5%

ϵ	θ	$e(m, w)$	$e(t, w)$	$e(s, w)$
0.05	0.000	0.83615	0.83615	0.69638
	0.250	0.84208	0.86082	0.70317
	0.500	0.86294	0.89467	0.72332
	1.000	0.97967	0.94991	0.79873
	1.500	1.23220	1.07476	0.90046
	2.000	1.61877	1.26326	0.98765
	2.500	2.08962	1.46462	1.02746
	3.000	2.59458	1.64439	1.02776

Table 16.2: Bahadur efficiencies of the mean test, t-test and the sign test with respect to the Wilcoxon test when the level of contamination is 10%

ϵ	θ	$e(m, w)$	$e(t, w)$	$e(s, w)$
0.10	0.000	0.72819	0.72819	0.72689
	0.250	0.73488	0.75380	0.73400
	0.500	0.75811	0.81043	0.75505
	1.000	0.87895	0.91614	0.83311
	1.500	1.12328	1.05802	0.93433
	2.000	1.47700	1.24380	1.01025
	2.500	1.67873	1.33772	1.02781
	3.000	2.32384	1.58321	1.02866

Table 16.3: Bahadur efficiencies of the mean test, t-test and the sign test with respect to the Wilcoxon test when the level of contamination is 25%

ϵ	θ	$e(m, w)$	$e(t, w)$	$e(s, w)$
0.25	0.000	0.61885	0.61885	0.82077
	0.250	0.62890	0.63528	0.82785
	0.500	0.65991	0.68437	0.84847
	1.000	0.79132	0.86165	0.91906
	1.500	1.01506	1.06753	0.99199
	2.000	1.30210	1.24270	1.02229
	2.500	1.61679	1.38110	1.00918
	3.000	1.94835	1.49314	0.98289

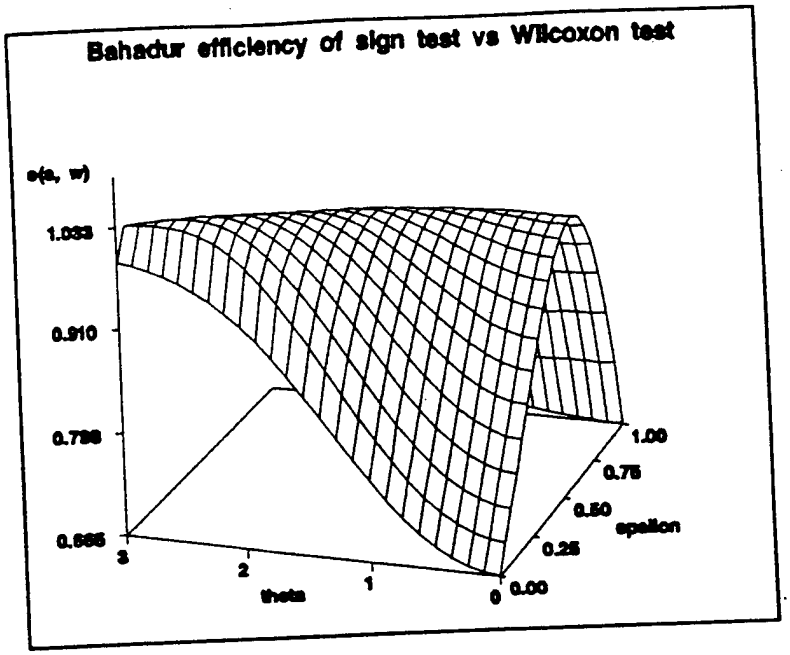

Figure 16.1: Bahadur efficiency of the mean test with respect to the Wilcoxon test

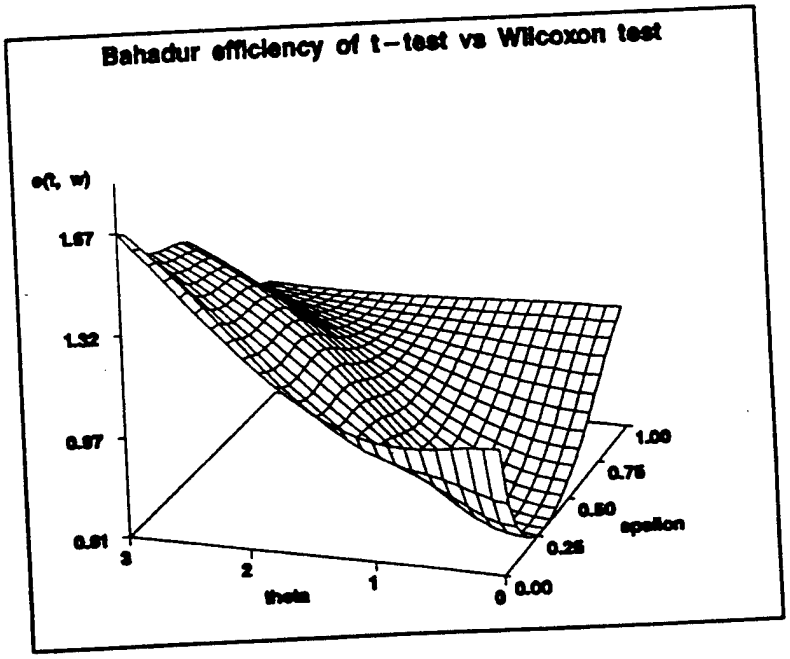

Figure 16.2: Bahadur efficiency of the t-test with respect to the Wilcoxon test

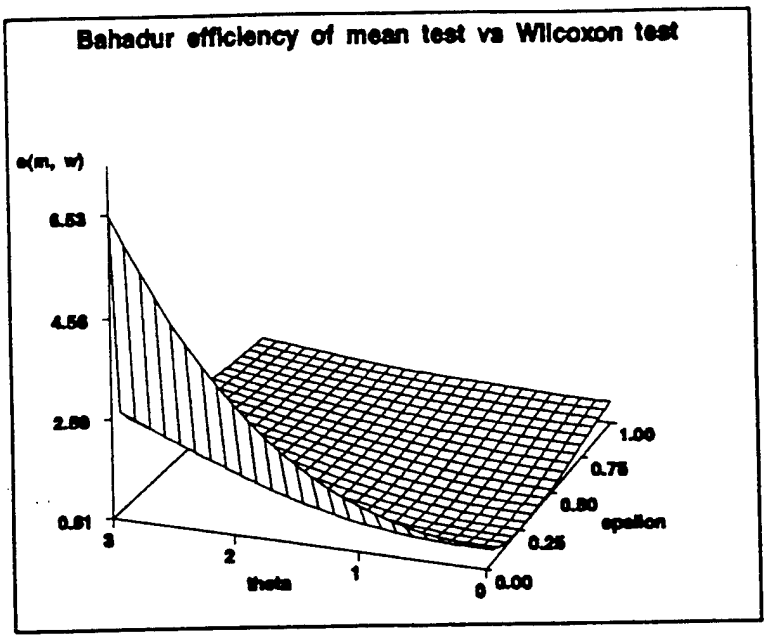

Figure 16.3: Bahadur efficiency of the sign test with respect to the Wilcoxon test

Reprinted from

journal of statistical planning and inference

Journal of Statistical Planning and Inference 63 (1997) 39–54

An alternative approach to the analysis of longitudinal data via generalized estimating equations

N. Rao Chaganty*

Department of Mathematics and Statistics, Old Dominion University, Norfolk, VA 23529-0077, USA

Received 9 April 1996; revised 13 September 1996

ELSEVIER

ELSEVIER

Journal of Statistical Planning and Inference 63 (1997) 39–54

journal of statistical planning and inference

An alternative approach to the analysis of longitudinal data via generalized estimating equations

N. Rao Chaganty*

Department of Mathematics and Statistics, Old Dominion University, Norfolk, VA 23529-0077, USA

Received 9 April 1996; revised 13 September 1996

Abstract

The generalized estimating equations (GEE) introduced by Liang and Zeger (*Biometrika* 73 (1986) 13–22) have been widely used over the past decade to analyze longitudinal data. The method uses a generalized quasi-score function estimate for the regression coefficients, and moment estimates for the correlation parameters. Recently, Crowder (*Biometrika* **82** (1995) 407–410) has pointed out some pitfalls with the estimation of the correlation parameters in the GEE method. In this paper we present a new method for estimating the correlation parameters which overcomes those pitfalls. For some commonly assumed correlation structures, we obtain unique feasible estimates for the correlation parameters. Large sample properties of our estimates are also established. © 1997 Elsevier Science B.V.

AMS classification: 62J12; 62F10; 62F12

Keywords: GEE; Longitudinal data; Positive definite; Quasi-likelihood; Repeated measures; Generalized least squares

1. Introduction

The statistical analysis of longitudinal data has been the topic of numerous statistical papers in recent years. Several books on the topic have also been published, for example Diggle et al. (1994), Jones (1993) and Lindsey (1993). Such data naturally occur when repeated observations are taken on individuals, or the data is taken on clusters or groups of subjects sharing similar characteristics.

In a landmark paper, Liang and Zeger (1986) introduced the generalized estimating equations (GEE) for analyzing longitudinal data. The setup and the method can be briefly described as follows. Let $Y_i=(y_{i1},\ldots,y_{it_i})'$ be a vector of repeated measurements taken on the ith subject; associated with each measurement y_{ij} is a vector of covariates $x_{ij}=(x_{ij1},\ldots,x_{ijp})'$, $1\leqslant j\leqslant t_i$, $1\leqslant i\leqslant m$. We will assume that the Y_i's are uncorrelated. We do not specify the joint distribution of the vector Y_i, but do make some

* Tel. +1 804 683 3897; fax: +1 804 683 3885; e-mail: nrc@math.odu.edu.

PII S0378-3758(96)00203-0

assumptions concerning the moments. Let $E(y_{ij})=\mu_{ij}$; $\text{var}(y_{ij})=\phi h(\mu_{ij})$, where $\phi>0$ may be a known constant or an unknown scale parameter. The variance function h is assumed to be a known function. We also assume that there is an invertible function g, known as the 'link function', such that $\mu_{ij}=g^{-1}(x_{ij}'\beta)$, where $\beta=(\beta_1,\ldots,\beta_p)'$ is a vector of regression coefficients. Our main parameter of interest is β. For simplicity, we will consider the balanced case in this paper; henceforth, we will set $t_i=t$ for all $1\leqslant i\leqslant m$. Extensions of our results to the unbalanced case and missing data situations will appear elsewhere.

The idea of Liang and Zeger (1986) is to model the dependence among the repeated measurements on the ith subject, in the form of a 'working correlation matrix' $R(\alpha)$, which is assumed to be a function of a vector of parameters $\alpha=(\alpha_1,\ldots,\alpha_q)'$. The covariance matrix of Y_i is then given by $\phi\Sigma_i$, where $\Sigma_i=A_i^{1/2}(\beta)R(\alpha)A_i^{1/2}(\beta)$ and $A_i(\beta)=\text{diag}(h(\mu_{i1}),h(\mu_{i2}),\ldots,h(\mu_{it}))$. The parameter α is considered to be a nuisance parameter. Let $\mathcal{S}$ be the subset of $\mathbb{R}^q$ such that $R(\alpha)$ is a positive-definite matrix for $\alpha\in\mathcal{S}$. In most examples the set $\mathcal{S}$ is an open convex subset of $\mathbb{R}^q$ and $R(\alpha)$ converges to a positive semi-definite matrix as α approaches the boundary of the set $\mathcal{S}$. Liang and Zeger (1986) suggest estimating β using a GEE and estimating α and ϕ using moment estimates via the current Pearsonian residuals. The estimate of β based on the GEE is essentially a multivariate analog of the quasi-score function estimate based on quasi-likelihood method. See Wedderburn (1974) and McCullagh (1983).

The GEE approach has several inherent pitfalls. Liang and Zeger (1986) have established consistency and asymptotic normality of their estimate of β as $m\to\infty$. Their proof depended on the use of $m^{1/2}$ consistent estimates for both α and ϕ. Using simple calculations, Crowder (1995) has demonstrated that there can be no general asymptotic theory supporting existence or consistency of the joint distribution of the estimates of β and α. He also used examples to show that the moment estimate of α might not fall in the set $\mathcal{S}$ of feasible values if the correlation structure is misspecified, thus crippling the whole estimation procedure. Prentice (1988) suggested another GEE for the estimate of α and established asymptotic normality for the joint distribution of his estimates of β and α. Prentice and Zhao (1991), extending the idea of Prentice (1988), introduced estimating equations in an ad hoc fashion for the covariance parameter estimation for a general multivariate response. There is no guarantee, however, the suggested estimates of the correlation parameters in either of those papers will fall within the set $\mathcal{S}$ of feasible values for small and even for moderately large samples.

In this paper we give a new approach to estimating the nuisance parameter α. Our method can be regarded as an extension of the method of (generalized) least squares, where we assume that the elements of the covariance matrix are functions of the regression parameters, moreover the off-diagonal elements are also functions of some unknown nuisance parameters. A partial minimization is then performed both with respect to the regression parameters as well as the unknown nuisance parameters.

In addition to yielding feasible estimates of α, our approach has several other advantages. The method of Prentice (1988) is computationally intensive, whereas in this paper we have closed-form expressions for our estimates of α for some correlation

structures. The method of Liang and Zeger (1986), for some correlation models, requires estimation of ϕ a priori to the estimation of β and α. Our estimates of β and α are independent of the value of ϕ. Unlike in Liang and Zeger (1986, Theorem 2), we will establish consistency and asymptotic normality of our estimate of β, without making any assumptions about the asymptotic properties of the estimates of α and ϕ.

The organization of this paper is as follows. In Section 2 we will give a motivation and derive an alternative set of estimating equations for β and α. The estimating equation for β is same as the GEE, whereas the estimating equation for α is new. The method of solving the estimating equations will be discussed in Section 3. In Section 4 the existence of a unique feasible estimate, $\hat{\alpha}$, for α will be established and a closed-form expression for $\hat{\alpha}$ for most of the commonly assumed correlation structures will be derived. In Section 5 we derive the large sample properties of our estimates, in particular, showing that $\hat{\beta}$ is consistent and $\hat{\alpha}$ is asymptotically biased. We will also prove that $\hat{\beta}$ and $\hat{\alpha}$ are jointly asymptotically normal and obtain expressions for the asymptotic covariances, and furthermore will show that the asymptotic distribution of $\hat{\alpha}$ does not depend on β. In Section 6 we present some simulation results, which show that for small samples, our estimate of β is highly efficient compared with the GEE estimate. Finally, the proofs are given in the appendix.

2. Estimating equations

This section outlines our new method of estimation of the unknown parameters β, α, and ϕ. For the longitudinal data setup described in Section 1, it is clear since we have no knowledge of the underlying distribution, that we should think of estimating the unknown parameters by the principle of (generalized) least squares. This requires minimizing the quadratic form

$$Q_\phi(\beta,\alpha) = \frac{1}{\phi}\sum_{i=1}^{m}(Y_i-\mu_i(\beta))'\Sigma_i^{-1}(Y_i-\mu_i(\beta))$$

$$= \frac{1}{\phi}\sum_{i=1}^{m}(Y_i-\mu_i(\beta))'A_i^{-1/2}(\beta)R^{-1}(\alpha)A_i^{-1/2}(\beta)(Y_i-\mu_i(\beta)). \tag{2.1}$$

Equating to zero the partial derivative with respect to α of (2.1) gives the first set of estimating equations:

$$\sum_{i=1}^{m} Z_i' \frac{\partial R^{-1}(\alpha)}{\partial \alpha_j} Z_i = 0, \quad 1 \leqslant j \leqslant q. \tag{2.2}$$

where $Z_i = A_i^{-1/2}(\beta)(Y_i - \mu_i(\beta))$, $1 \leqslant i \leqslant m$. Strictly speaking the estimating equation for β should now be obtained by differentiating (2.1) with respect to β. However, we would like to avoid certain complications that arise with the differentiation, caused by β appearing both in the mean vector $\mu_i(\beta)$ and the variance matrix $A_i(\beta)$. And importantly we would like to get an estimating equation which yields an unbiased

estimate for β in some cases where α is known. Finally, we would like our estimate of β to coincide with the maximum likelihood estimate in cases where the observations y_{ij}'s are a random sample from an exponential family of distributions. To derive an estimating equation for β which satisfies the three aforementioned requirements, we will treat the quadratic form (2.1) as a function of three variables

$$Q(\beta,\beta^*,\alpha)=\sum_{i=1}^{m}(Y_i-\mu_i(\beta))'A_i^{-1/2}(\beta^*)R^{-1}(\alpha)A_i^{-1/2}(\beta^*)(Y_i-\mu_i(\beta)). \tag{2.3}$$

Carrying out the differentiation of (2.3) with respect to β, then substituting $\beta^*=\beta$, gives the estimating equation

$$\sum_{i=1}^{m}D_i'(\beta)A_i^{-1/2}(\beta)R^{-1}(\alpha)Z_i=0, \tag{2.4}$$

where $D_i(\beta)=\partial\mu_i/\partial\beta'$.

Solving Eq. (2.4) for β amounts to searching the minimum values of (2.3) along cross sections perpendicular to the β^* axis, and choosing the value that falls on the 45° line with respect to the (β,β^*) axis. But the principle of (generalized) least squares requires finding the infimum of (2.3) along the 45° line with respect to the (β,β^*) axis. In general these two methods of minimization do not yield the same estimate for β, though they do coincide if the global infimum of (2.3) with respect to (β,β^*) happens to fall on the 45° line. Eq. (2.4) is exactly the equation proposed by Liang and Zeger (1986) to estimate β, and can also be derived using the principle of quasi-likelihood.

Our method of estimating the parameters is to solve the estimating Eqs. (2.2) and (2.4) simultaneously for β and α to obtain estimates $\widehat{\beta}$ and $\widehat{\alpha}$. A step by step recursive algorithm for solving the equations, based on a Fisher scoring method similar to the one proposed by Liang and Zeger (1986), is given in Section 3. From the definition of $\widehat{\beta}$ and $\widehat{\alpha}$ we have

$$Q(\beta,\beta,\alpha)\geqslant Q(\widehat{\beta},\beta,\widehat{\alpha}) \quad \text{for all } \beta,\alpha \tag{2.5}$$

where the function Q is defined in (2.3). Since the estimates do not fully conform to the principle of (generalized) least squares, it is reasonable to call our estimates $\widehat{\beta}$, and $\widehat{\alpha}$, 'quasi-least squares estimates' of β and α, respectively.

Suppose that ϕ is an unknown scale parameter; it is playing the same role as σ^2 of ordinary least squares (OLS) theory. See Rao (1973, p. 227). In OLS, σ^2 is estimated using the mean residual sum of squares, and the same approach here says to estimate ϕ by

$$\widehat{\phi}=\frac{1}{mt}\sum_{i=1}^{m}\widehat{Z}_i'\widehat{Z}_i \tag{2.6}$$

where $\widehat{Z}_i=A_i^{-1/2}(\widehat{\beta})(Y_i-\mu_i(\widehat{\beta}))$. If a bias-corrected estimate is preferable, we can use $\widehat{\phi}_b=mt\widehat{\phi}/(mt-p)$, instead.

3. Iterative procedure for the estimates

We will now study methods of solving Eqs. (2.2) and (2.4) for β and α. Let

$$f(\alpha)=\sum_{i=1}^{m} Z_i' R^{-1}(\alpha) Z_i, \tag{3.1}$$

where $Z_i = A_i^{-1/2}(\beta)(Y_i - \mu_i(\beta))$, for $1 \leqslant i \leqslant m$. Eq. (2.2) can be rewritten as

$$\frac{\partial f(\alpha)}{\partial \alpha_j} = \sum_{i=1}^{m} Z_i' \frac{\partial R^{-1}(\alpha)}{\partial \alpha_j} Z_i = -\sum_{i=1}^{m} Z_i' R^{-1}(\alpha) \frac{\partial R(\alpha)}{\partial \alpha_j} R^{-1}(\alpha) Z_i = 0, \quad 1 \leqslant j \leqslant q. \tag{3.2}$$

For many commonly employed correlation structures where $R^{-1}(\alpha)$ is readily available, Eq. (3.2) can be solved explicitly for α in terms of Z_i's. In other cases an alternate way of solving Eq. (2.2) is to use the spectral decomposition

$$R(\alpha) = P(\alpha) \Lambda(\alpha) P'(\alpha),$$

where $P(\alpha)$ is an orthogonal matrix of eigenvectors and $\Lambda(\alpha) = \operatorname{diag}(\lambda_k(\alpha))$ is a diagonal matrix consisting of the eigenvalues of $R(\alpha)$. We will see later for some commonly employed correlation structures $R(\alpha)$, the matrix of eigenvectors $P(\alpha) = P$ does not depend on α, but only the eigenvalues $\Lambda(\alpha)$ depend on α. In this case we can rewrite Eq. (3.1) as

$$\begin{aligned} f(\alpha) = \sum_{i=1}^{m} Z_i' P \Lambda^{-1}(\alpha) P' Z_i &= \sum_{i=1}^{m} W_i' \Lambda^{-1}(\alpha) W_i \\ &= \sum_{i=1}^{m} \sum_{k=1}^{t} \frac{w_{ik}^2}{\lambda_k(\alpha)} \\ &= \sum_{k=1}^{t} \left(\sum_{i=1}^{m} w_{ik}^2 \Big/ \lambda_k(\alpha) \right), \end{aligned} \tag{3.3}$$

where $W_i = P' Z_i = (w_{ik})$. Differentiating (3.3) with respect to α we get the following set of estimating equations:

$$\frac{\partial f(\alpha)}{\partial \alpha_j} = \frac{\partial}{\partial \alpha_j} \left\{ \sum_{k=1}^{t} \left(\sum_{i=1}^{m} w_{ik}^2 \Big/ \lambda_k(\alpha) \right) \right\} = 0, \quad 1 \leqslant j \leqslant q, \tag{3.4}$$

which can be used instead of (3.2) to get an estimate of α.

An iterative method for obtaining the estimates $\hat{\beta}$, $\hat{\alpha}$ of β and α respectively, can be described as follows:

Step 1: Choose an initial value $\tilde{\beta}$ for β.

Step 2: Compute $\tilde{A}_i = A_i(\tilde{\beta})$, $\tilde{\mu}_i = \mu_i(\tilde{\beta})$, $\tilde{Z}_i = \tilde{A}_i^{-1/2}(Y_i - \tilde{\mu}_i)$ and $\tilde{D}_i = D_i(\tilde{\beta})$, $1 \leqslant i \leqslant m$.

Step 3: Solve for $\tilde{\alpha}$ using either Eq. (3.2) or Eq. (3.4). Compute $\tilde{R} = R(\tilde{\alpha})$ and $\tilde{\Sigma}_i = \tilde{A}_i^{-1/2} \tilde{R} \tilde{A}_i^{-1/2}$, $1 \leqslant i \leqslant m$.

Step 4: Update the value of β as

$$\widehat{\beta}=\widetilde{\beta}+\left\{\sum_{i=1}^{m}\widetilde{D}_i'\widetilde{\Sigma}_i^{-1}\widetilde{D}_i\right\}^{-1}\left\{\sum_{i=1}^{m}\widetilde{D}_i'\widetilde{\Sigma}_i^{-1}(Y_i-\widetilde{\mu}_i)\right\}.$$

Step 5: Stop the process if $\widehat{\beta}\simeq\widetilde{\beta}$ and take $\widehat{\beta}$ as an estimate of β. The estimate of α is given by $\widehat{\alpha}=\widetilde{\alpha}$. Otherwise, repeat Steps 2–4 replacing $\widetilde{\beta}$ by $\widehat{\beta}$.

In the next section we will show that the estimate $\widetilde{\alpha}$ in Step 3 falls within the set $\mathcal{S}$ of feasible values at every step of the iteration for commonly assumed correlation structures.

4. Special correlation structures

We will now discuss solutions for Eq. (3.2) for commonly assumed correlation structures. In some of the examples, we will also obtain unique feasible, closed form solutions for the estimate of α. In general, that there is a unique solution $\widetilde{\alpha}\in\mathcal{S}$ for the equations given by (3.2) can be shown as follows. Let us denote the first and second order partial derivatives of $R(\alpha)$ by the matrices

$$\frac{\partial R(\alpha)}{\partial\alpha}=\operatorname{diag}\left(\frac{\partial R(\alpha)}{\partial\alpha_j}\right),\qquad\frac{\partial^2R(\alpha)}{\partial\alpha^2}=\left(\frac{\partial^2R(\alpha)}{\partial\alpha_j\partial\alpha_{j'}}\right)$$

respectively, both of order $qt\times qt$. We will use the symbol $\otimes$ to denote the Kronecker product between two matrices. It is easy to verify that the matrix of second-order partial derivatives of $f(\alpha)$ defined in (3.1) can be written as

$$\nabla^2f(\alpha)=2\sum_{i=1}^{m}\left\{(I_q\otimes Z_i'R^{-1})\frac{\partial R(\alpha)}{\partial\alpha}(ee'\otimes R^{-1})\frac{\partial R(\alpha)}{\partial\alpha}(I_q\otimes R^{-1}Z_i)\right.$$
$$\left.-(I_q\otimes Z_i'R^{-1})\frac{\partial^2R(\alpha)}{\partial\alpha^2}(I_q\otimes R^{-1}Z_i)\right\},$$

where $R^{-1}=R^{-1}(\alpha)$ and I_q is the identity matrix of order q and e is a $t\times1$ column vector of ones. For several correlation structures, the elements of the correlation matrix $R(\alpha)$ are linear functions of α and we have $\partial^2R(\alpha)/\partial\alpha^2=0$. It is therefore easy to verify that $\nabla^2f(\alpha)$ is a positive definite matrix for $\alpha\in\mathcal{S}$, for all m, or in some cases for $m\geqslant t$. Hence $f(\alpha)$ is a strictly convex function. Furthermore, $f(\alpha)\to\infty$ as α approaches the boundary of $\mathcal{S}$. It thus has a unique minimum at $\widetilde{\alpha}\in\mathcal{S}$, where $\widetilde{\alpha}$ is such that $\nabla f(\widetilde{\alpha})=0$. Note that in Examples 4.1–4.3 below $q=1$ and $\alpha_1=\rho$.

Example 4.1. Suppose that the observations on each subject are equicorrelated with correlation ρ. The correlation matrix equals $R(\rho)=(1-\rho)I_t+\rho ee'$, where $\rho\in\mathcal{S}=(-1/(t-1),\,1)$. For this correlation structure, it is well known that

$$R^{-1}(\rho)=\left\{\frac{1}{(1-\rho)}I_t-\frac{\rho}{(1-\rho)(1+(t-1)\rho)}ee'\right\}.$$

Thus in this case (3.2) reduces to

$$\left\{\sum_{i=1}^{m} Z_i' Z_i - \frac{1+(t-1)\rho^2}{(1+(t-1)\rho)^2}\sum_{i=1}^{m}(e'Z_i)^2\right\}=0. \tag{4.1}$$

Let $\overline{Z}_i=(e'Z_i)/t$ and $S_i^2=\{Z_i'(I_t-ee'/t)Z_i\}/(t-1)$ be the mean and variance of the components of the vector Z_i. If we let $d=t\,(\sum_{i=1}^m \overline{Z}_i^2)/(\sum_{i=1}^m S_i^2)$ then Eq. (4.1) can be written as

$$\left\{1+\frac{t\rho}{(1-\rho)}\right\}^2=d. \tag{4.2}$$

The value of $\rho\in\mathcal{S}$ satisfying (4.2) is given by

$$\tilde{\rho}=\frac{d^{1/2}-1}{d^{1/2}+(t-1)}. \tag{4.3}$$

We can use the above value of $\tilde{\rho}$ in Step 3 of the iterative process for this correlation structure. In this example the method of Liang and Zeger (1986) requires the estimation of ϕ a priori to the estimation of ρ, whereas our estimate $\tilde{\rho}$ given by (4.3) is independent of ϕ.

Example 4.2. Let the correlation matrix $R(\rho)$ be a tridiagonal matrix, with 1 on the diagonal and ρ on the upper and lower diagonals. This is equivalent to the one-dependent model. The eigenvalues and eigenvectors of $R(\rho)$ are given by

$$\lambda_k(\rho)=1+2\rho\cos\{k\pi/(t+1)\},\quad 1\leqslant k\leqslant t$$

and

$$x_k=(\sin\{k\pi/(t+1)\},\ldots,\sin\{tk\pi/(t+1)\})',\quad 1\leqslant k\leqslant t,$$

respectively. We can verify that $R(\rho)$ is positive definite if and only if $\rho\in\mathcal{S}=(\rho_1,\rho_t)$, where $\rho_k=-1/(2\cos\{k\pi/(t+1)\})$. Clearly, in this example the eigenvectors do not depend on ρ. Since x_k's are not orthonormal we can construct, using Gram–Schmidt orthogonalization, a set of orthonormal eigenvectors $\{p_k,\ 1\leqslant k\leqslant t\}$ from x_k's. Let $W_i=P'Z_i=(w_{ik})$, where the kth column of P is p_k. In this case we can verify that Eq. (3.4) reduces to

$$f'(\rho)=\frac{\mathrm{d}}{\mathrm{d}\rho}\left\{\sum_{k=1}^{t}\frac{\sum_{i=1}^m w_{ik}^2}{\lambda_k(\rho)}\right\}=\frac{\mathrm{d}}{\mathrm{d}\rho}\left\{\sum_{k=1}^{t}\frac{\sum_{i=1}^m w_{ik}^2}{(1+2\rho\cos\{k\pi/(t+1)\})}\right\}$$

$$=-\left\{\sum_{k=1}^{t}\frac{2\cos\{k\pi/(t+1)\}\sum_{i=1}^m w_{ik}^2}{(1+2\rho\cos\{k\pi/(t+1)\})^2}\right\}$$

$$=0.$$

It is easy to check that $f'(\rho_1)=-\infty$, $f'(\rho_t)=\infty$ and $f'(\rho)$ is continuous on the interval (ρ_1,ρ_t). Therefore, there exists a $\tilde{\rho}$ such that $f'(\tilde{\rho})=0$. This establishes

the existence of a solution for the above equation. The value of $\tilde{\rho}$ can be computed numerically and can be used in Step 3 of the iterative process. We can also handle l-dependent ($l>1$) structure in a similar manner, though the expressions for the eigenvalues and eigenvectors are not as simple as the case $l=1$. Unlike the method of Liang and Zeger (1986) no estimate of ϕ is required to get $\tilde{\rho}$ in this example.

Example 4.3. Suppose the correlation matrix $R(\rho)=(\rho^{|i-j|})$, where $\rho \in \mathcal{S}=(-1,1)$. This structure is the well known first-order autoregressive (AR(1)) structure. Note that

$$R^{-1}(\rho)=\frac{1}{(1-\rho^2)}\{I_t+\rho^2 C_2-\rho C_1\},$$

where $C_2=\operatorname{diag}(0,1,\ldots,1,0)$ and C_1 is a tridiagonal matrix with 0 on the diagonal and 1 on the upper and lower diagonals. Thus

$$f(\rho)=\sum_{i=1}^{m} Z_i' R^{-1}(\rho) Z_i$$

$$=\frac{1}{(1-\rho^2)}\left\{\sum_{i=1}^{m} Z_i' Z_i+\rho^2\sum_{i=1}^{m} Z_i' C_2 Z_i-\rho\sum_{i=1}^{m} Z_i' C_1 Z_i\right\}.$$

Equating to zero the derivative of $f(\rho)$ we get

$$a_m\rho^2-2b_m\rho+a_m=0, \tag{4.4}$$

where $a_m=\sum_{i=1}^{m} Z_i' C_1 Z_i$ and $b_m=\sum_{i=1}^{m} Z_i'(I_t+C_2)Z_i$. Note that the elements of $R(\rho)$ are not linear functions of ρ in this example. We will show in Appendix A that there is a unique root for Eq. (4.4) in the interval $\mathcal{S}=(-1,1)$ and is given by

$$\tilde{\rho}=\frac{b_m-\{b_m^2-a_m^2\}^{1/2}}{a_m}. \tag{4.5}$$

The value of $\tilde{\rho}$ can be used in Step 3 of the iterative process. In this example, if we use the method of Liang and Zeger (1986), an estimate of ϕ must be computed in the determination of the estimate of β, whereas our method does not require estimation of ϕ prior to the estimation of β.

Example 4.4. We now consider the case where the correlation matrix R is totally unspecified. To get an estimate of R, we need to

$$\min_R \sum_{i=1}^{m} Z_i' R^{-1} Z_i = \min_R \operatorname{tr}(ZR^{-1}), \tag{4.6}$$

where $Z=\sum_{i=1}^{m} Z_i Z_i'$. Let us assume that $m \geqslant t$; which is a reasonable assumption, considering the fact that we have $t(t-1)/2$ unknown correlation parameters. The matrix Z is positive definite in this case. It has been shown by Whittle (1958, p. 234, Lemma 3) that there exists a unique, positive-definite correlation matrix $\tilde{R}$ where the minimum (4.6) is attained. See also Olkin and Pratt (1958). The correlation matrix $\tilde{R}$, can be

obtained by solving the equation

$$Z = \tilde{R}\Delta\tilde{R}, \tag{4.7}$$

where Δ is a diagonal matrix of positive elements. It follows from the results of Olkin and Pratt (1958, p. 231) that the solution to Eq. (4.7) is given by

$$\tilde{R} = \Delta^{-1/2}(\Delta^{1/2}Z\Delta^{1/2})^{1/2}\Delta^{-1/2} \tag{4.8}$$

and the diagonal matrix Δ satisfies the fixed point equation

$$\Delta = \text{diag}(\Delta^{1/2}Z\Delta^{1/2})^{1/2}. \tag{4.9}$$

For a given Z, the diagonal matrix Δ satisfying (4.9) can be obtained recursively starting with a trial value Δ_0 and computing $\Delta_k = \text{diag}(\Delta_{k-1}^{1/2}Z\Delta_{k-1}^{1/2})^{1/2}$ at the kth step. The proof that this fixed point iteration scheme converges to the unique solution of Eq. (4.9) and related results will appear elsewhere. The estimate $\tilde{R}$ given by Eq. (4.8) can be used in Step 3 of the iterative process.

5. Large sample properties

In this section, we will study the large sample properties of the quasi-least square estimates $\hat{\beta}$ and $\hat{\alpha}$ defined in Section 2. In particular, we will show that $\hat{\beta}_m$ is consistent, whereas $\hat{\alpha}_m$ is asymptotically biased as $m \to \infty$; the subscript m emphasizes the dependence of the estimates on m. Theorem 5.1 below shows that the joint distribution of $(\hat{\beta}_m, \hat{\alpha}_m)$ is asymptotically normal. We will introduce some notation before stating the main theorem of this section. Let $\bar{R}$ be the true correlation matrix. Recall that $R(\alpha)$ is the working correlation matrix.

Assume that $\Phi = E(Z_i \otimes Z_i Z_i')$ and $\Psi = E(Z_i Z_i' \otimes Z_i Z_i')$ are finite, where the expectation is taken under the true correlation structure $\bar{R}$. Let $\lambda = (\beta, \alpha, \phi)'$ and $\theta = (\beta, \alpha)'$. Define

$$\mathscr{I}_{i11}(\theta) = \{D_i'(\beta)A_i^{-1/2}(\beta)R^{-1}(\alpha)A_i^{-1/2}(\beta)D_i(\beta)\}_{p\times p}, \tag{5.1}$$

$$\mathscr{V}_{i11}(\theta) = \{D_i'(\beta)A_i^{-1/2}(\beta)R^{-1}(\alpha)\bar{R}R^{-1}(\alpha)A_i^{-1/2}(\beta)D_i(\beta)\}_{p\times p}, \tag{5.2}$$

$$\mathscr{V}_{i12}(\lambda) = [\text{tr}[\{(e_j'D_i'(\beta)A_i^{-1/2}(\beta)R^{-1}(\alpha)) \otimes B_k\}\Phi]]_{p\times q}, \tag{5.3}$$

where $B_k = \partial R^{-1}(\alpha)/\partial\alpha_k$ and e_j is $p \times 1$ column vector with one at the jth row and zero elsewhere. Note that if the working correlation is indeed the true correlation then $\mathscr{V}_{i11}(\theta) = \mathscr{I}_{i11}(\theta)$. The following three quantities are useful to describe the asymptotic distribution of $\hat{\alpha}_m$:

$$a(\alpha) = [\text{tr}\{B_j\bar{R}\}]_{q\times 1},$$

$$\mathscr{I}_{22}(\alpha) = \left[\text{tr}\left\{\frac{\partial^2 R^{-1}(\alpha)}{\partial\alpha_j\,\partial\alpha_k}\bar{R}\right\}\right]_{q\times q},$$

$$\mathscr{V}_{22}(\lambda) = [\,\text{tr}\{(B_j \otimes B_k)\Psi\}/\phi^2 - \text{tr}\{B_j\bar{R}\}\text{tr}\{B_k\bar{R}\}]_{q\times q}. \tag{5.4}$$

Define

$$\mathscr{I}_i(\lambda)=\begin{bmatrix} 2\mathscr{I}_{i11}(\theta) & 0 \\ 0 & \phi\mathscr{I}_{22}(\alpha)\end{bmatrix}, \qquad \mathscr{V}_i(\lambda)=\begin{bmatrix} 4\phi\mathscr{V}_{i11}(\theta) & \mathscr{V}_{i12}(\lambda) \\ \mathscr{V}'_{i12}(\lambda) & \phi^2\mathscr{V}_{22}(\lambda)\end{bmatrix}. \tag{5.5}$$

Assume that

$$\frac{1}{m}\sum_{i=1}^{m}\mathscr{I}_i(\lambda)\to\begin{bmatrix} 2\mathscr{I}_{11}(\theta) & 0 \\ 0 & \phi\mathscr{I}_{22}(\alpha)\end{bmatrix}=\mathscr{I}(\lambda) \quad \text{(say)} \tag{5.6}$$

and

$$\frac{1}{m}\sum_{i=1}^{m}\mathscr{V}_i(\lambda)\to\begin{bmatrix} 4\phi\mathscr{V}_{11}(\theta) & \mathscr{V}_{12}(\lambda) \\ \mathscr{V}'_{12}(\lambda) & \phi^2\mathscr{V}_{22}(\lambda)\end{bmatrix}=\mathscr{V}(\lambda) \quad \text{(say)} \tag{5.7}$$

as $m\to\infty$.

We are now in a position to state the main theorem of this section. The regularity conditions needed to establish Theorem 5.1 are same as the conditions that we would normally use in a multivariate central limit theorem for independent but not necessarily identically distributed random vectors. In fact, conditions (5.6), (5.7) are similar to the condition on the covariance matrices that is in the multivariate central limit theorem, Theorem B of Serfling (1981, p. 30).

Theorem 5.1. *Let $\lambda=(\beta,\alpha,\phi)'$ be fixed. Let $\theta=(\beta,\alpha)'$ and $\widehat{\theta}_m=(\widehat{\beta}_m,\widehat{\alpha}_m)'$ be the solution to Eqs.* (2.2) *and* (2.4). *Suppose that conditions* (5.6) *and* (5.7) *hold. Then $\widehat{\beta}_m$ is a consistent estimate of β, whereas $\widehat{\alpha}_m$ is asymptotically biased. Further,*

$$(\widehat{\theta}_m-\theta) \text{ is AN}\left(\mathscr{I}^{-1}(\lambda)\mu(\lambda), \frac{\mathscr{I}^{-1}(\lambda)\mathscr{V}(\lambda)\mathscr{I}^{-1}(\lambda)}{m}\right) \tag{5.8}$$

where $\mu(\lambda)=(0,\phi a'(\alpha))'$ and $\mathscr{I}(\lambda)$, $\mathscr{V}(\lambda)$, $a(\alpha)$ are defined in (5.6), (5.7) *and* (5.4).

Proof of Theorem 5.1 is given in Appendix B. It is easy to check that (5.8) implies that $\widehat{\beta}_m$ and $\widehat{\alpha}_m$ are asymptotically correlated and

$$\widehat{\beta}_m \text{ is AN}\left(\beta, \frac{\phi\mathscr{I}_{11}^{-1}(\theta)\mathscr{V}_{11}(\theta)\mathscr{I}_{11}^{-1}(\theta)}{m}\right),$$

$$\widehat{\alpha}_m \text{ is AN}\left(\alpha+\mathscr{I}_{22}^{-1}(\alpha)a(\alpha), \frac{\mathscr{I}_{22}^{-1}(\alpha)\mathscr{V}_{22}(\lambda)\mathscr{I}_{22}^{-1}(\alpha)}{m}\right).$$

We can also easily verify that $\widehat{\phi}_m$ given in (2.6) is a consistent estimate of ϕ using the fact that $\widehat{\beta}_m$ is a consistent estimate of β, even if the working correlation is misspecified. Note that if the working correlation is correctly specified, that is, $R(\alpha)=\overline{R}$, then $\mathscr{V}_{11}(\theta)=\mathscr{I}_{11}(\theta)$ and the asymptotic covariance of $\widehat{\beta}_m$ reduces to $\phi\mathscr{I}_{11}^{-1}(\theta)/m$. Since in practice the true correlation $\overline{R}$ is unknown, we can assume that the working correlation is correctly specified. An estimate of the covariance matrix of $\widehat{\beta}_m$ is then obtained by replacing the parameters α and β in (5.1) with their estimates, giving

us

$$\widehat{\text{cov}}_1(\widehat{\beta}_m) = \widehat{\phi}_m \left\{ \sum_{i=1}^{m} D_i'(\widehat{\beta}_m)\widehat{\Sigma}_i^{-1} D_i(\widehat{\beta}_m) \right\}^{-1} \tag{5.9}$$

where $\widehat{\Sigma}_i = A_i^{1/2}(\widehat{\beta}_m)R(\widehat{\alpha}_m)A_i^{1/2}(\widehat{\beta}_m)$ and $\widehat{\phi}_m$ is given by (2.6). Alternatively, following Liang and Zeger (1986), Royall (1986), we can estimate the covariance matrix of $\widehat{\beta}_m$ using a model-robust, sandwich-type variance estimator given by

$$\widehat{\text{cov}}_2(\widehat{\beta}_m) = \left\{ \sum_{i=1}^{m} D_i'(\widehat{\beta}_m)\widehat{\Sigma}_i^{-1} D_i(\widehat{\beta}_m) \right\}^{-1} \left\{ \sum_{i=1}^{m} D_i'(\widehat{\beta}_m)\widehat{\Sigma}_i^{-1} \widehat{\text{cov}}(Y_i)\widehat{\Sigma}_i^{-1} D_i(\widehat{\beta}_m) \right\}$$

$$\times \left\{ \sum_{i=1}^{m} D_i'(\widehat{\beta}_m)\widehat{\Sigma}_i^{-1} D_i(\widehat{\beta}_m) \right\}^{-1} \tag{5.10}$$

where $\widehat{\text{cov}}(Y_i) = \widehat{U}_i\widehat{U}_i'$, $U_i = Y_i - \mu_i(\widehat{\beta}_m)$. The estimate (5.9) or (5.10) can be used to construct confidence intervals for linear functions of β.

Remark 5.1. It is interesting to note that the asymptotic distribution of $\widehat{\alpha}_m$ does not depend on β, unlike the asymptotic distribution of $\widehat{\beta}_m$, which depends on all the parameters β, α, ϕ and $\overline{R}$. Also, the asymptotic bias of $\widehat{\alpha}_m$ depends only on α and $\overline{R}$ but not on ϕ.

Remark 5.2. In the case where the distribution of Z_i's is correctly specified and it is a multivariate normal distribution with mean 0 and covariance matrix $\phi R(\alpha)$ and if $\partial^2 R(\alpha)/\partial\alpha^2 = 0$, then we can show that $\mathscr{V}_{22}(\lambda) = \mathscr{I}_{22}(\alpha)$. Thus in this case we have

$$\widehat{\alpha}_m \text{ is AN}\left(\alpha + \mathscr{I}_{22}^{-1}(\alpha)a(\alpha), \frac{\mathscr{I}_{22}^{-1}(\alpha)}{m}\right). \tag{5.11}$$

Remark 5.3. For some working correlation structures, simulation results have shown that it is possible to reduce the bias of $\widehat{\alpha}_m$ using the jackknife, bias-reducing technique. Also, since the asymptotic covariance of $\widehat{\alpha}_m$ depends on the third and fourth moments of the y_{ij}'s, it is perhaps best to use the nonparametric method, bootstrap, to estimate the covariance of $\widehat{\alpha}_m$. See Efron (1982) for an excellent introduction to the jackknife and the bootstrap methods. On the other hand, in data analysis problems where α is also an important parameter, the GEE method is preferable, since it uses a consistent estimate of α, provided of course the GEE estimate of α falls within the set of feasible values.

6. Simulation results for small samples

In this section we will show, using Monte Carlo simulations, that the relative efficiency of the quasi-least squares regression parameter estimates can be very high

for small samples when compared to the estimates obtained using the GEE method. To make a fair comparison between the two methods, we consider an example where a totally unspecified working correlation structure is appropriate, since in this case the GEE method also yields feasible estimates for the correlation parameters. We will first fit a model using the GEE method to a real-life data and then make a comparison between the two methods using simulated data from the fitted model.

Consider the data in Table 3.10 of Rencher (1995, p. 92). The data consists of blood glucose levels on three occasions for 52 women. The variable y represents a fasting glucose measurement and the covariate x is the glucose measurement one hour after sugar intake. We have fit a simple linear regression model between y and x for the data using the GEE method with identity link function ($g(u)=u$), and totally unspecified working correlation structure. The regression line is estimated to be

$$y = 61.364 + 0.1098x. \tag{6.1}$$

The estimates of the correlation matrix between the three repeated measurements, and of the scale parameter, are given by

$$R_0 = \begin{pmatrix} 1 & 0.1971 & -0.0122 \\ 0.1971 & 1 & 0.2081 \\ -0.0122 & 0.2081 & 1 \end{pmatrix}, \qquad \phi_0 = 76.65. \tag{6.2}$$

The regression coefficients in (6.1) were highly significant. We will use the model (6.1) and (6.2), which describes the relationship between the three repeated measurements on y and x, to compare the quasi-least squares and the GEE methods.

To make the comparison between the two methods, we have simulated 1000 replications of samples of three ($t=3$) repeated measurements on the variable y, using the x values in Table 3.10 of Rencher (1995, p. 92) on m women for $m=5, 15, 52$. The simulations were performed using the values of the parameters in (6.1) and (6.2) and a Guassian distribution for the errors. We then fit the true model for each replication of the simulated data using the quasi-least squares and the GEE methods. For quasi-least squares we have used the fixed-point iteration scheme described in Example 4.4 to estimate the correlation matrix in the iterative process for obtaining the regression parameter estimates. Mean square errors (MSE) of the estimates of the intercept and the slope were computed using the 1000 replications.

Table 1 gives the relative efficiencies of the quasi-least squares estimates with respect to the GEE estimates of the regression parameters for various values of m. The relative efficiency is defined as the ratio of the MSE computed from the GEE method to that of the MSE of the quasi-least squares method. We can see from Table 1 that the relative efficiency is very high for $m=5$, being more than three for both the intercept and the slope. Further, the relative efficiency is decreasing as m increases. But note that even for moderately large samples ($m=52$, the size of the original sample in Rencher (1995, p. 92)) the relative efficiency of the quasi-least squares is more than 1. Therefore, the quasi-least squares approach is preferable not only for small samples, but for moderately large samples as well.

Table 1
Relative efficiencies of the quasi-least squares regression parameters estimates with respect to the GEE estimates.

m	Intercept	Slope
5	3.387	3.183
15	1.219	1.236
52	1.057	1.062

Table 2
Bias/(standard error) of the estimates of the correlation parameters. The numbers above diagonal correspond to the quasi-least squares estimates; numbers below diagonal are for the GEE estimates.

$m=5$			$m=15$			$m=52$		
*	0.1693 (0.2850)	0.0135 (0.2657)	*	0.1184 (0.1350)	0.0112 (0.1395)	*	0.1030 (0.0709)	0.0051 (0.0743)
0.0137 (0.5901)	*	0.1575 (0.2708)	0.0377 (0.2715)	*	0.1267 (0.1369)	0.0126 (0.1390)	*	0.1105 (0.0704)
0.0144 (0.5680)	0.0755 (0.5639)	*	0.0250 (0.2817)	0.0461 (0.2722)	*	0.0128 (0.1493)	0.0171 (0.1378)	*

The biases and the standard errors of the estimates of the correlation parameters are contained in Table 2. As expected, the quasi-least squares estimates have more bias than those from the GEE method. On the other hand, the quasi-least squares estimates of the correlation parameters have smaller standard errors and therefore are more stable than the GEE estimates.

Acknowledgements

I am grateful to Ms. Justine Shults for stimulating my interest in the GEE. The Editor's and a referee's comments on an earlier version of this paper have been very helpful. I would also like to thank Dr. J.P. Morgan for editing, Dr. D.N. Naik and Dr. A.K. Vaish for useful discussions. This research was partially supported by the U.S. Army research office grant number DAAH04-96-1-0070. The United States Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon.

Appendix A. Proof of unique feasible root in Example 4.3

We can verify that in Example 4.3, the roots of the quadratic equation (4.4) are real if and only if

$$\left\{\sum_{i=1}^{m} Z_i'(I_t + C_2)Z_i\right\}^2 \geqslant \left\{\sum_{i=1}^{m} Z_i' C_1 Z_i\right\}^2$$

$$\Leftrightarrow 1 \geqslant \max_{z \neq 0} \left[\frac{z'(I_m \otimes C_1)z}{z'\{I_m \otimes (I_t + C_2)\}z} \right]^2 \tag{A1}$$

$$\Leftrightarrow 1 \geqslant |\lambda_{\max}\{(I_m \otimes C_1)(I_m \otimes (I_t + C_2))^{-1}\}|$$

$$\Leftrightarrow 1 \geqslant |\lambda_{\max}\{C_1(I_t + C_2)^{-1}\}|,$$

where $z' = (Z_1', \ldots, Z_m')$ and $\lambda_{\max}(A)$ denotes the maximum eigenvalue of A. The matrix $C_1(I_t + C_2)^{-1}$ is similar to a symmetric tridiagonal matrix. Using the strum sequence property of tridiagonal matrices, we can verify that all the eigenvalues of $C_1(I_t + C_2)^{-1}$ fall in the interval $[-1, 1]$. Therefore (A1) holds and we have established that

$$\left| \sum_{i=1}^{m} Z_i' C_1 Z_i \right| \leqslant \sum_{i=1}^{m} Z_i'(I_t + C_2) Z_i \tag{A2}$$

for all Z_i's. It is easy to verify, using the inequality (A2), that the root of Eq. (4.4) that falls in the interval $(-1, 1)$ is given by (4.5) almost surely.

Appendix B. Proof of Theorem 5.1

Fix $\theta = (\beta, \alpha)'$. Let $v_i(\beta^*, \theta) = Z_i'(\beta^*, \beta) R^{-1}(\alpha) Z_i(\beta^*, \beta)$, where $Z_i(\beta^*, \beta) = A_i^{-1/2}(\beta^*)$ $(Y_i - \mu_i(\beta))$. Now, for $\|t\| \leqslant K$, $0 < K < \infty$, under standard regularity conditions, considering a Taylor series expansion around θ we can write

$$\xi_m(t) = \sum_{i=1}^{m} \{v_i(\beta^*, \theta + t/m^{1/2}) - v_i(\beta^*, \theta)\}$$

$$= \frac{1}{m^{1/2}} \sum_{i=1}^{m} t' \nabla v_i(\beta^*, \theta) + \frac{1}{2m} \sum_{i=1}^{m} t' \nabla^2 v_i(\beta^*, \theta^*) t \tag{B1}$$

where θ^* is a point on the line joining θ and $\theta + t/m^{1/2}$. The above expansion is true for any (β^*, β, α). In particular for $\beta^* = \beta$, we can write (B1) as

$$\xi_m(t) = \frac{1}{m^{1/2}} \sum_{i=1}^{m} t' \nabla v_i(\theta) + \frac{1}{2m} \sum_{i=1}^{m} t' \nabla^2 v_i(\theta^*) t. \tag{B2}$$

Now, if we let

$$\eta_m(t) = \frac{1}{m} \sum_{i=1}^{m} \{\nabla^2 v_i(\theta^*) - \nabla^2 v_i(\theta)\},$$

then (B2) can be rewritten as

$$\xi_m(t) = \frac{1}{m^{1/2}} \sum_{i=1}^{m} t' \nabla v_i(\theta) + \frac{1}{2m} \sum_{i=1}^{m} t' \nabla^2 v_i(\theta) t + \frac{t' \eta_m(t) t}{2}$$

$$= \frac{1}{m^{1/2}} \sum_{i=1}^{m} t' \nabla v_i(\theta) + \frac{1}{2m} \sum_{i=1}^{m} t' \mathscr{I}_i(\lambda) t$$

$$+ \frac{1}{2m} \sum_{i=1}^{m} t' \{\nabla^2 v_i(\theta) - \mathscr{I}_i(\lambda)\} t + \frac{t' \eta_m(t) t}{2}, \tag{B3}$$

where $E\{\nabla^2 v_i(\theta)\} = \mathscr{I}_i(\lambda)$. Under the assumption (5.6) it follows from the weak law of large numbers,

$$\frac{1}{2m}\sum_{i=1}^{m} t'\{\nabla^2 v_i(\theta) - \mathscr{I}_i(\lambda)\}t = o_p(1).$$

Using an argument similar to the one found in Sen and Singer (1993, p. 207), we can show that $\sup_{\{t:\,\|t\|\leqslant K\}} |\eta_m(t)|$ tends to zero almost surely as $m \to \infty$. Thus uniformly for $\|t\| \leqslant K$ we have

$$\xi_m(t) = \frac{1}{m^{1/2}}\sum_{i=1}^{m} t'\nabla v_i(\theta) + \frac{1}{2m}\sum_{i=1}^{m} t'\mathscr{I}_i(\lambda)t + o_p(1). \tag{B4}$$

Disregarding the $o_p(1)$ term and minimizing the right-hand side of (B4) with respect to t, we get the point of minimum as

$$\hat{t}_m = \left\{\frac{1}{m}\sum_{i=1}^{m}\mathscr{I}_i(\lambda)\right\}^{-1}\left\{\frac{1}{m^{1/2}}\sum_{i=1}^{m}\nabla v_i(\theta)\right\}. \tag{B5}$$

From the definition of $\xi_m(t)$ we can conclude that $\hat{t}_m$ will also correspond closely to the quasi-least squares estimate of θ, which is attained at $\hat{\theta}_m$. Therefore, we have

$$\hat{\theta}_m = \theta + \frac{\hat{t}_m}{m^{1/2}} + o_p\left(\frac{1}{m^{1/2}}\right). \tag{B6}$$

It is easy to verify that

$$E\{\nabla v_i(\theta)\} = \mu(\lambda) \quad \text{and} \quad \operatorname{cov}\{\nabla v_i(\theta)\} = \mathscr{V}_i(\lambda). \tag{B7}$$

From (B5)–(B7) and the weak law of large numbers, we get

$$\hat{\theta}_m \to \theta + \mathscr{I}^{-1}(\lambda)\mu(\lambda) \tag{B8}$$

in probability, as $m \to \infty$. It is easy to check from (B8) that $\hat{\beta}_m$ is consistent and $\hat{\alpha}_m$ is asymptotically biased. Under the assumptions (5.6), (5.7), from (B7), (B6), (B5), the multivariate central limit theorem and Slutsky's theorem, we can conclude that

$$(\hat{\theta}_m - \theta) = \frac{\hat{t}_m}{m^{1/2}} + o_p\left(\frac{1}{m^{1/2}}\right) \text{ is } \mathrm{AN}\left(\mathscr{I}^{-1}(\lambda)\mu(\lambda), \frac{\mathscr{I}^{-1}(\lambda)\mathscr{V}(\lambda)\mathscr{I}^{-1}(\lambda)}{m}\right). \tag{B9}$$

This completes the proof of the theorem.

References

Crowder, M. (1995). On the use of a working correlation matrix in using generalised linear models for repeated measures. *Biometrika* **82**, 407–410.

Diggle, P., K-Y. Liang and S.L. Zeger (1994). *Analysis of Longitudinal Data.* Oxford University Press, Oxford.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. CBMS-NSF Regional conference series in Applied Mathematics. SIAM, Philadelphia.

Jones, R.M. (1993). *Longitudinal Data with Serial Correlation: A State Space Approach*. Chapman and Hall, London.

Liang, K-Y. and S.L. Zeger (1986). Longitudinal data analysis using generalised linear models. *Biometrika* **73**, 13–22.

Lindsey, J.K. (1993). *Models for Repeated Measurements*. Oxford University Press, Oxford.

McCullagh, P. (1983). Quasi-likelihood functions. *Ann. Statist.* **11**, 59–67.

Prentice, R.L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44**, 1033–1048.

Prentice, R.L. and L.P. Zhao (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics* **47**, 825–839.

Olkin, I. and J.W. Pratt (1958). A multivariate Tchebycheff inequality. *Ann. Math. Statist.* **29**, 226–234.

Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed. Wiley, New York.

Rencher, A.C. (1995). *Methods of Multivariate Analysis*. Wiley, New York.

Royall, R.M. (1986). Model robust inference using maximum likelihood estimators. *Internat. Statist. Rev.* **54**, 221–226.

Sen, P.K. and J.M. Singer (1993). *Large Sample Methods in Statistics: An Introduction with Applications*. Chapman & Hall, New York.

Serfling, R.J. (1981). *Approximation Theorems of Mathematical Statistics*, Wiley, New York.

Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalised linear models, and the Gauss–Newton method. *Biometrika* **61**, 439–447.

Whittle, P. (1958). A multivariate generalization of Tchebichev's inequality. *Quart. J. Math. Oxford Ser.* **9**, 232–240.

BIOMETRICS 54, 1622–1630
December 1998

Analysis of Serially Correlated Data Using Quasi-Least Squares

Justine Shults*

Center for Pediatric Research, Eastern Virginia Medical School, Norfolk, Virginia 23510-1001, U.S.A.

and

N. Rao Chaganty

Department of Mathematics and Statistics, Old Dominion University, Norfolk, Virginia 23529-0077, U.S.A.

SUMMARY

Quasi-least squares (QLS), a marginal statistical approach via generalized estimating equations that is described in the balanced data setting by Chaganty (1997, *Journal of Statistical Planning and Inference* **63**, 39–54), allows for application of a wide range of working correlation structures when analyzing serially correlated data. We extend the application of QLS to serially correlated, unequally spaced, and unbalanced data using three useful working correlation models: the first-order autoregressive (AR(1)), the Markov, and the generalized Markov structure described by Núñez-Anton and Woodworth (1994, *Biometrics* **50**, 445–456). We compare QLS and the original formulation of the generalized estimating equation approach (GEE) for these structures, demonstrating that (i) infeasibility of the GEE correlation parameter estimates can be a problem, (ii) it is difficult to obtain consistent moment estimates of the correlation parameters for the generalized Markov structure, and (iii) the use of QLS can lead to reduced mean square error of the estimate of the regression parameter for small samples of moderately correlated data. To choose between alternative correlation models, we propose a criterion that is based on the principle of generalized least squares. Finally, data for which the generalized Markov structure is appropriate are analyzed to demonstrate the use of QLS in selecting a suitable working correlation structure and identifying important covariates.

1. Introduction

In this paper, we apply a statistical method based on the generalized estimating equation approach of Liang and Zeger (1986) to the analysis of longitudinal data that may be difficult to analyze using other established methods. We consider repeated measures data collected by taking measurements of an outcome variable and associated covariates on each of a group of independent subjects. Our primary data analysis goal is to identify important covariates and to explain their effect on the marginal mean of the outcome variable while also accounting for the correlation among observations on each subject.

Accomplishing this research objective may be difficult due to certain conditions that are typical in longitudinal studies. The timing and total number of measurements taken may vary from subject to subject so that the data may be unbalanced and unequally spaced. The outcome variable may not be normally distributed. The intrasubject correlation may be described using a time-dependent pattern. For example, the correlation between two measurements may decrease as they become more

* *Corresponding author's email address:* jshults@chkd.com

Key words: Cholesky decomposition; GEE; Longitudinal data; Positive definite matrix; Quasi-least squares; Serial correlation.

highly separated in time, or two measurements separated by a fixed distance in time may be more highly correlated if they are collected later during the study rather than earlier.

To describe the marginal mean of the outcome variable as a function of the covariates, one approach is to use the method of generalized estimating equations (GEE), first proposed by Liang and Zeger (1986). The method of GEE specifies a generalized linear model for the outcome variable and models the association among observations on each subject via a working correlation structure. Estimation proceeds by alternating between solving a generalized estimating equation for the regression parameter and consistent moment estimation of the correlation parameter. Many recent publications discuss extensions and applications of the GEE approach, especially for correlated binary data. In particular, Prentice (1988) and Zhao and Prentice (1990) developed generalizations of GEE that Liang, Zeger, and Qaqish (1992) refer to as GEE1 and GEE2, respectively. Desmond (1997) gives a good description of GEE, GEE1, and GEE2.

One widely accepted property of the method of GEE is that, if the correlation structure is misspecified, the estimates of the regression parameters will nevertheless remain consistent. There has been some controversy, however, regarding the effect of misspecification on the efficiency of these estimates. [For a discussion of this topic, see papers by McDonald (1993), Zhao, Prentice, and Self (1992), Fitzmaurice and Laird (1993), and Fitzmaurice (1995).] In any case, it is intuitively reasonable that careful modeling of the correlation structure leads to improved estimation of the regression and the correlation parameters.

Modeling the correlation structure of the outcomes on each subject comprises (i) identifying reasonable correlation structures for the data under consideration, (ii) implementing these structures in an analysis, and (iii) choosing among the final sets of estimates associated with each of the different structures. Depending on our initial identification of reasonable working correlation structures, we may be limited in carrying out steps (ii) and (iii) using the method of GEE. For some correlation structures, implementing (ii) is difficult, either because the final GEE estimates of their parameters are infeasible or because consistent moment estimates of their parameters are not easily obtained. In this paper, we consider three successively generalized spatial correlation structures that are applicable to serially correlated data—the AR(1), the Markov, and the generalized Markov structure that was described by Núñez-Anton and Woodworth (1994). For the AR(1) and Markov structures, GEE may yield infeasible final estimates of the correlation parameters. For the generalized Markov structure, consistent moment estimates of the parameters are not easily obtained so that GEE is not easily applied for this structure. Carrying out (iii) using GEE may also be difficult because the method does not provide a simple criterion for correlation model selection.

In contrast to the original formulation of GEE, QLS does provide a simple basis for nonasymptotic comparison of different correlation structures. We suggest a criterion for correlation model selection that is based on the principle of generalized least squares. The QLS approach also allows for consideration of the AR(1), Markov, and generalized Markov correlation structures and, for a continuous outcome variable, the final QLS estimates of the parameters in these structures will be feasible. The goal of this paper is to demonstrate that, when compared with the original formulation of GEE, QLS can improve our ability to model the correlation in our data. We also conduct simulations to show that QLS can lead to more efficient estimation of the regression parameters. Comparisons between QLS, GEE1, and GEE2 for correlated binary data are planned as the subject of future research.

Organization of this paper is as follows. In Section 2, we establish notation, give a description of the method of quasi-least squares, and propose a criterion for correlation model selection. In Section 3, we discuss the AR(1), Markov, and generalized Markov correlation structures, give an interpretation of their correlation parameters, and discuss implementation of the method of QLS for each structure. In Section 4, we describe simulations that compare QLS with GEE and a data analysis that demonstrates the use of QLS in choosing an appropriate working correlation structure and identifying important covariates.

2. The Method of Quasi-Least Squares

2.1 *Notation and Assumptions*

We consider data comprising vectors $Y_i' = (y_{i1}, y_{i2}, \ldots, y_{in_i})$ of measurements taken on subject i at times $T_i' = (t_{i1}, t_{i2}, \ldots, t_{in_i})$, $0 < t_{i1} < t_{i2} < \cdots < t_{in_i}$. Associated with each measurement y_{ij} is a vector of covariates $x_{ij}' = (x_{ij1}, x_{ij2}, \ldots, x_{ijp})$; $1 \le j \le n_i$, $1 \le i \le m$. We assume that any variability in the spacing or number of observations collected on each subject is either the result of the study design or of a process that is independent of the observed and unobserved data, i.e., missing values are missing at random. We do not assume a distributional form for the outcome

variable, only that $E(y_{ij}) = u_{ij}$ and $\text{var}(y_{ij}) = \tau_j \nu(u_{ij})$, where $\tau_j > 0$ is either a known constant or an unknown parameter. The regression equation $u_{ij} = g^{-1}(x'_{ij}\beta)$ relates the marginal mean of the outcome variable with covariates measured on each subject, where $\beta' = (\beta_1, \beta_2, \ldots, \beta_p)$ is a vector of unknown regression coefficients and g is an invertible link function.

We assume that observations taken on different subjects are independent. Taken on one subject, they are correlated. The covariance matrix V_i of observations on subject i satisfies $V_i = (\Upsilon_i A_i)^{1/2} R_i(\rho)(A_i \Upsilon_i)^{1/2}$, where $A_i = \text{diag}(\nu(u_{i1}), \ldots, \nu(u_{in_i}))$, $\Upsilon_i = \text{diag}(\tau_1, \ldots, \tau_{n_i})$, $R_i(\rho)$ is a working correlation matrix, and $\rho' = (\rho_1, \rho_2, \ldots, \rho_s)$ is a vector of unknown parameters. We consider ρ to be a nuisance parameter; its estimation is carried out primarily to aid estimation of β. Let $U'_i = (u_{i1}, \ldots, u_{in_i})$ and $Z_i(\beta) = A_i^{-1/2}(Y_i - U_i) = (z_{i1}, \ldots, z_{in_i})'$. We refer to the quadratic form

$$Q(\rho, \beta) = \sum_{i=1}^{m} Z'_i(\beta) R_i^{-1}(\rho) Z_i(\beta) \tag{2.1}$$

as the generalized error sum of squares.

2.2 *A Description of the Methods*

Here we describe the method of QLS and make a brief comparison with the original formulation of GEE. [For a more detailed description in the balanced data setting, see Chaganty (1997).] QLS uses a partial derivative of the generalized error sum of squares (2.1) to derive the following estimating equation for β:

$$\sum_{i=1}^{m} D'_i A_i^{-1/2} R_i^{-1}(\rho) Z_i(\beta) = 0. \tag{2.2}$$

The estimating equation for ρ is obtained by differentiating (2.1) with respect to ρ, yielding

$$\sum_{i=1}^{m} Z'_i(\beta) \frac{dR_i^{-1}(\rho)}{d\rho} Z_i(\beta) = 0. \tag{2.3}$$

To obtain QLS estimates $(\hat{\rho}, \hat{\beta})$ for (ρ, β), we select a starting value $\tilde{\beta}$ for β (or a starting value $\tilde{\rho}$ for ρ) and then iterate between solving (2.3) for ρ (the rho step) and solving (2.2) for β (the beta step) until the estimates of β converge. GEE also uses estimating equation (2.2) for β and alternates between estimation of ρ and β, though it requires the use of $m^{1/2}$ consistent estimates of ρ. In practice, moment estimates of ρ that are based on the current values of the standardized residuals (z_{ij}) are often used. If τ_j is an unknown parameter, it can be consistently estimated using $\hat{\tau}_j = (1/m)\sum_{i=1}^{m} \hat{z}_{ij}^2$, where $\hat{z}_{ij}$ is z_{ij} evaluated at $\hat{\beta}$. If $\tau_j = \tau$ for all j, we use the consistent estimate $\hat{\tau} = (1/n)\sum_{i=1}^{m}\sum_{j=1}^{n_i} \hat{z}_{ij}^2$, where $n = \sum_{i=1}^{m} n_i$.

2.3 *Choosing a Correlation Structure*

To choose among competing structures, Diggle, Liang, and Zeger (1994, p. 145) suggest fitting different correlation models and then comparing the corresponding final estimates of the regression parameter and their standard errors. "If they differ substantially, a more careful treatment of the covariance model may be necessary." There are two drawbacks to this approach. First, it relies on asymptotic standard errors. As pointed out by Lindsey (1993, p. 68), decisions based on asymptotic standard errors may be inappropriate for small samples. Second, if more careful modeling of the correlation structure seems necessary, there is ambiguity as to how one should then proceed with the analysis.

In contrast to the method of GEE, QLS provides a basis for nonasymptotic comparison of different correlation structures. Since the method estimates ρ by minimizing (2.1) with respect to ρ, given several alternatives, a natural choice of correlation structure is the structure that minimizes the generalized error sum of squares, with an adjustment for the total number of parameters in the model. If, e.g., we are analyzing data in which the correlation among measurements on each subject is expected to decrease with increased separation in time, we might reasonably apply the structures considered in Section 3. For a given set of covariates and a specified link function and mean variance relationship, we might obtain QLS estimates of (ρ, β) for each structure, choosing as our final correlation model the structure that corresponds to the minimum value of the adjusted residual generalized sum of squares $\hat{Q}_a = Q(\hat{\rho}, \hat{\beta})/(n - p - q)$, where $n = \sum_{i=1}^{m} n_i$. Our criterion generalizes to correlated data the least-squares approach of minimizing the residual sum of squares.

It is analogous to Akaike's Information Criterion (AIC) and Schwarz's Bayesian Criterion (SBC), two criteria commonly used for covariance model selection in analyses of normally distributed data (see Littell et al., 1996, p. 101).

Another approach that can be used to compare the fit of two correlation models is construction of an empirical semivariogram (Diggle et al., 1994, p. 82) or the two graphical techniques, draftman's display and parallel axis plots, suggested by Dawson, Gennings, and Carter (1997).

3. Three Useful Spatial Correlation Models

Here we discuss three useful correlation structures, interpret their parameters, and discuss implementation of QLS for each structure.

3.1 *The Correlation Structures*

In what follows, $\rho = (\rho_1, \rho_2) = (\alpha, \lambda)$. Let $R_i(\alpha, \lambda) = [r^i_{jk}]$, where $r^i_{jk} = \alpha^{e_{ij+1}+\cdots+e_{ik}}$ for $k > j$, $r^i_{jj} = 1$, $r^i_{jk} = r^i_{kj}$ for $k < j$, and e_{ij} is a function of (α, λ). In this paper, we consider three special cases of $R_i(\alpha, \lambda)$: the AR(1), for which $e_{ij} = 1$ for all i and j; the Markov, for which $e_{ij} = t_{ij} - t_{ij-1}$; and the generalized Markov, for which

$$e_{ij} = \begin{cases} \frac{t^\lambda_{ij} - t^\lambda_{ij-1}}{\lambda} & \text{if } \lambda \neq 0;\ 0 < t_{ij-1} < t_{ij} \\ \ln\left(\frac{t_{ij}}{t_{ij-1}}\right) & \text{if } \lambda = 0;\ 0 < t_{ij-1} < t_{ij}. \end{cases} \tag{3.1}$$

We can easily verify that $R_i(\alpha, \lambda)$ has a unique Cholesky decomposition $\Gamma_i(\alpha, \lambda)\Gamma'_i(\alpha, \lambda)$, where $\Gamma_i(\alpha, \lambda)$ is a lower triangular matrix (see the Appendix). Since the kth diagonal element of $\Gamma_i(\alpha, \lambda)$ is $(1 - \alpha^{2e_{ik}})^{1/2}$ and $R_i(\alpha, \lambda)$ is positive definite if and only if all the diagonal elements of $\Gamma_i(\alpha, \lambda)$ are positive, feasible values of α are those values for which $1 - \alpha^{2e_{ik}}$ is defined and positive for all k and i.

Bounds on α for each structure are as follows. The parameter $\alpha \in (-1, 1)$ for the AR(1) structure and also for the Markov structure if the e_{ik} are integer valued. To allow α to take on negative values for the Markov structure, if necessary, we could change the time scale so that the e_{ik} are integers or are suitably defined so that all values of $1 - \alpha^{2e_{ik}}$ are defined and positive. This may be problematic, however. Suppose that e_{ij+1} is odd, so that $\text{corr}(y_{ij}, y_{ij+1})$ may be either positive or negative. Changing the time scale to make e_{ij+1} even is equivalent to assuming that $\text{corr}(y_{ij}, y_{ij+1})$ is positive. Since our basic assumptions regarding the model should be invariant to choice of time scale, we use the Markov structure only when we expect the intrasubject correlation to be positive, which is the case in most biological applications. We thus restrict α to the interval $(0, 1)$. For the generalized Markov, $e_{ij} > 0$ for any fixed $\lambda \in (-\infty, \infty)$ and we restrict α to $(0, 1)$, as for the Markov structure.

The parameters (α, λ) have a useful interpretation for longitudinal data analysis. For the AR(1) or Markov structure, the correlation between measurements on one subject decreases with increasing difference in order or timing of measurements, respectively. The Markov structure is appropriate for unequally spaced observations and may be used as an alternative to imputation of missing values. The generalized Markov structure introduces an additional parameter λ to the Markov structure that greatly increases its flexibility. We first note that the generalized model allows for accelerated, or decelerated, decay in the correlation between measurements for a fixed value of α since $(t^\lambda_{ik} - t^\lambda_{ij})/\lambda$ increases towards ∞ as $\lambda \to \infty$ and decreases towards zero as $\lambda \to -\infty$. (Here we have assumed that (i) $t_{ij} > 1$, which can be achieved through reparameterization in the time scale, if necessary, and (ii) that $t_{ij} < t_{ik}$, for all i, and $k > j$.) This is useful because one potential difficulty in applying the Markov structure is that it may force the correlations between measurements to decrease too rapidly with increasing separation in time. Other researchers, including Muñoz et al. (1992), also used parameters to dampen the correlation in the Markov structure. The generalized Markov model extends the Markov structure so that the correlation between measurements is not just a function of their separation in time but also of their time of occurrence in the study. This is because $\lim_{\lambda \to 0}(t^\lambda_{ik} - t^\lambda_{ij})/\lambda = \ln(t_{ik}/t_{ij}) = \ln(1 + w/t_{ij})$, where $t_{ik} = w + t_{ij}$. Since, for fixed w, $\lim_{t_{ij} \to \infty} \ln(1 + w/t_{ij}) = 0$, for values of λ that are small in absolute value, responses in the outcome variable that are separated by w time units will be more highly correlated if they are observed later in the study than if they are observed earlier. This generalization will be useful if we are dealing with outcome variables, such as growth in humans, that become more highly correlated over time.

3.2 *Implementing the Method of Quasi-Least Squares*

To simplify programming the beta step of QLS, we use the Cholesky decomposition $R_i^{-1}(\alpha,\lambda) = L_i(\alpha,\lambda)L_i'(\alpha,\lambda)$ (see the Appendix) and an approach similar to that described in Lindsey (1993). Using current estimates $\tilde{\alpha}$ and $\tilde{\lambda}$, we calculate $T_i' = L_i'(\tilde{\alpha},\tilde{\lambda})\tilde{A}_i^{-1/2}\tilde{e}_i$ and $S_i' = L_i'(\tilde{\alpha},\tilde{\lambda})\tilde{A}_i^{-1/2}\tilde{D}_i$ for all i, where $\tilde{e}_i = Y_i - \tilde{U}_i$. We then regress $T = (T_1', T', \ldots, T_m')'$ on $S = (S_1', S_2', \ldots, S_m')'$ to obtain an adjustment that is added to our previous estimate of β. The estimating process ends when this adjustment is approximately zero.

To implement the ρ step, we again use the Cholesky decomposition to reexpress (2.1) as

$$Q(\alpha,\lambda,\beta) = \sum_{i=1}^{m}\sum_{j=2}^{n_i} \frac{z_{ij}^2 - 2\alpha^{e_{ij}} z_{ij}z_{ij-1} + z_{ij-1}^2}{1-\alpha^{2e_{ij}}} - \sum_{i:n_i>2}\sum_{j=2}^{n_i-1} z_{ij}^2. \tag{3.2}$$

For the AR(1) structure, we used differentiation and simple arithmetic to obtain the following unique point of minimum of (3.2) in the interval $(-1,1)$:

$$\tilde{\alpha} = \frac{\sum_{i=1}^{m}\sum_{j=2}^{n_i}(z_{ij}^2 + z_{ij-1}^2) - \sqrt{\sum_{i=1}^{m}\sum_{j=2}^{n_i}(z_{ij}-z_{ij-1})^2 \sum_{i=1}^{m}\sum_{j=2}^{n_i}(z_{ij}+z_{ij-1})^2}}{2\sum_{i=1}^{m}\sum_{j=2}^{n_i} z_{ij}z_{ij-1}}. \tag{3.3}$$

For the Markov correlation structure, we used a modified Newton–Raphson method to minimize (3.2) with respect to α over $(0,1)$. For the generalized Markov structure, we wrote a grid search program. All programs were written in STATA.

4. Comparison with the Method of GEE

Here we use simulations to compare the methods of GEE and QLS when the true and working correlation structures are both AR(1) (Section 4.2), Markov (Section 4.2), and generalized Markov (Section 4.3).

4.1 *The Model for the Simulations*

We consider a data set collected according to a two-treatment cross-over design because incorrectly assuming that the observations are uncorrelated in this setting can lead to a severe loss of efficiency in estimation of β (see Diggle et al., 1994, pp. 60–61). The model for our simulations is $Y_i = X_i\beta + \epsilon_i$, where

$$X_i' = \begin{pmatrix} 1 & 1 & 1 \\ x_{i1} & x_{i2} & x_{i3} \end{pmatrix}$$

and $\beta' = (\beta_0, \beta_1)$; $i = 1,2,\ldots,8$. The treatment sequences (x_{i1}, x_{i2}, x_{i3}) comprise all distinct permutations of zeros and ones for $i = 1,2,\ldots,8$. For the AR(1) structure, the measurements are equally spaced. For the Markov and generalized Markov structures, the vector of timings (t_{i1}, t_{i2}, t_{i3}) is given by $(2,7,8)$ for $i = 1,2,3,4$, $(2,3,8)$ for $i = 5,6,7$, and $(2,5,9)$ for $i = 8$. We allow the timings to vary between subjects because, although many study protocols call for a common set of measurement times, in practice, this goal is not often achieved. This lack of a common set of timings means that, assuming a common unstructured correlation matrix for all subjects, as is often done in practice, may not be appropriate. We assume constant variance, i.e., $\tau_j = \tau$ for $j = 1,2,3$. Correlation in the data is induced by ϵ_i, which is assumed to be multivariate normal with mean zero and covariance τR_i. We set $\tau = 4$ and $(\beta_0, \beta_1) = (120, -12.88)$. The correlation matrix R_i has AR(1) structure in Section 4.2, Markov structure in Section 4.2, and generalized Markov structure in Section 4.3.

4.2 *Comparisons for the* AR(1) *and Markov Structures*

For the AR(1) correlation structure, we made our comparisons using the closed-form QLS estimate of α and the following GEE estimate of α that is used in the "SAS Macro for Longitudinal Data Analysis" (Groemping, 1994):

$$\hat{\alpha}_G = \frac{(n-p)}{(n-m-p)} \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i-1} z_{ij}z_{ij+1}}{\sum_{i=1}^{m}\sum_{j=1}^{n_i} z_{ij}^2}, \tag{4.1}$$

where $n = \Sigma_{i=1}^{m} n_i$. For the Markov correlation structure, Liang and Zeger (1986, example 4) suggest an *ad hoc* estimate of α. They first note that $\mathrm{E}(z_{ij}z_{ik}) = \tau\alpha^{|t_{ij}-t_{ik}|}$. Substituting the current estimate $\tilde{z}_{ij}\tilde{z}_{ik}$ for $\mathrm{E}(z_{ij}z_{ik})$ and then taking logarithms yields $\ln(\tilde{z}_{ij}\tilde{z}_{ik}) \approx \ln(\tau) + \ln(\alpha)|t_{ij} - t_{ik}|$. A natural estimate of $\ln(\alpha)$ is then given by the slope of the regression line of $\ln(\tilde{z}_{ij}\tilde{z}_{ik})$ on

$|t_{ij} - t_{ik}|$. (When programming the method, we regressed $\ln(|\tilde{z}_{ij}\tilde{z}_{ik}|)$ on $|t_{ij} - t_{ik}|$ since $\tilde{z}_{ij}\tilde{z}_{ik}$ may take on negative values.) Infeasibility is a problem for this *ad hoc* estimate, so we used a modified estimate that was not so often infeasible during simulation runs. Consider the set $\{d_1 < d_2 < \cdots < d_l\}$ of distinctive values of spacings between any two measurements on one subject. Let s_w be the number of pairs (t_{ij}, t_{ik}) such that $|t_{ij} - t_{ik}| = d_w$. Since $\mathrm{E}(z_{ij}z_{ik}) = \tau\alpha^{d_w}$ if $|t_{ij} - t_{ik}| = d_w$, we estimate $\ln(\alpha)$ by the slope of the regression line of $\ln(H_w)$ on d_w, where $H_w = (1/s_w)\Sigma_{\{(t_{ij},t_{ik}):|t_{ij}-t_{ik}|=d_w\}}\,\tilde{z}_{ij}\tilde{z}_{ik}$.

We compared QLS and GEE for the Markov and AR(1) structures with regard to infeasibility and efficiency. Our simulations demonstrate the problem that GEE may have regarding infeasibility of its correlation parameter estimates. When the true and working structures are AR(1) and the data are equally spaced, approximately 10% of simulation runs yielded infeasible final GEE estimates of α. For the Markov structure and unequally spaced data, infeasibility was a greater problem. Although H_w is a superior estimate to $\tilde{z}_{ij}\tilde{z}_{ik}$, it may not estimate $\mathrm{E}(z_{ij}z_{ik})$ precisely for small samples, so that the slope of the regression line used to construct the Markov GEE estimates may be close to zero, resulting in an estimate of α that is close to one and maybe greater than one. In our simulations, the final GEE Markov correlation parameter estimates were positively biased for all values of α and were often infeasible. In some simulation runs, over 30% of the final GEE estimates of α were infeasible. Because GEE is usually implemented using moment estimates of the correlation parameters, feasibility of these estimates is not guaranteed. For *ad hoc* estimates, such as the Markov GEE estimate considered here, the likelihood of obtaining an infeasible correlation parameter estimate may be high. Since QLS will always yield feasible estimates for continuous outcome variables, QLS might prove useful in providing a feasible correlation parameter estimate should the method of GEE fail to do so.

Table 1 contains the ratio of the mean square error of the QLS regression parameter estimates to the mean square error of the GEE estimates when the true and working correlation structures are both AR(1) or both Markov. Simulation runs that yielded GEE estimates of α that were infeasible were not used in the comparison and the efficacy of GEE is thus overstated. Table 1 shows that QLS estimates β more efficiently than GEE when the intrasubject correlation is small to moderate ($\alpha \leq 0.5$), which is the case in most biological applications. For $\alpha > 0.5$, GEE estimates β more precisely than QLS, though it is important to bear in mind that, for all values of α, GEE may yield an infeasible final estimate of the correlation parameter. The relative performance of the two methods for larger values of α is probably due to properties of the moment estimates used by GEE to estimate α for the AR(1) and Markov correlation structures, which had lower mean square error for values of $\alpha \approx 1$. Simulations conducted by the authors also confirmed what Diggle et al. (1994, p. 60–61) observed: incorrectly assuming that the outcomes are uncorrelated in this setting leads to inefficiency in estimation of β, especially as the intrasubject correlation increases in value. Simulations were also performed for sample sizes 16, 32, 64, and 128. Even for the larger samples, QLS outperformed GEE in terms of mean square error for $\alpha \leq 0.5$. However, the gain in performance decreased with increasing m.

4.3 *Comparisons for the Generalized Markov Structure*

To demonstrate the difficulties involved, we attempt to extend Liang and Zeger's (1986) *ad hoc* approach to estimation of α for the generalized Markov structure. For simplicity, assume that each of m subjects has measurement times that are a subset of a common set of measurement times $\{t_1 < t_2 < \cdots < t_N\}$. Now consider the set of spacings $\{\eta_{jk}(\lambda) = (t_k^{\lambda} - t_j^{\lambda})/\lambda;\ 1 \leq j < k \leq N\}$.

Table 1

Efficiency of QLS regression parameter estimate with respect to the GEE estimate for the AR(1) *and Markov correlation structures*

	Group parameter		Constant parameter	
α	AR(1)	Markov	AR(1)	Markov
0.1	1.12	1.75	1.07	1.29
0.3	1.07	1.42	1.03	1.14
0.5	0.98	1.11	0.99	1.04
0.7	0.84	0.87	0.96	0.96
0.9	0.63	0.71	0.96	0.96

Table 2
MSEs of the regression parameter estimates obtained using QLS with three correlation structures

	Group parameter			Constant parameter		
λ	Identity	Markov	Generalized Markov	Identity	Markov	Generalized Markov
−5	2.61	0.09	0.05	2.54	1.90	1.89
−3	2.63	0.31	0.31	2.34	1.74	1.74
1	2.61	2.13	2.10	1.62	1.47	1.46
3	2.61	2.65	2.65	1.31	1.30	1.32

Let $n_{ijk} = 1$ if subject i has measurement times t_j and t_k and let $n_{ijk} = 0$ otherwise. Let $s_{jk} = \Sigma_{i=1}^{m} n_{ijk}$ and $H_{jk} = (1/s_{jk}) \Sigma_{\{i:n_{ijk}=1\}} \tilde{z}_{ij} \tilde{z}_{ik}$. Since $\mathrm{E}(z_{ij} z_{ik}) = \tau \alpha^{\eta_{jk}(\lambda)}$, we can estimate (α, λ) using nonlinear regression between H_{jk} and $\tau \alpha^{\eta_{jk}(\lambda)}$. Clearly, this procedure does not yield simple feasible and consistent estimates for the correlation parameters. The generalized Markov structure is thus not easily applied using GEE.

We conducted simulations according to the model described in Section 4.1, in which the actual correlation structure is generalized Markov. Table 2 contains the mean square error (MSE) of the QLS estimates of the regression parameters for $\alpha = 0.6$ and for various values of λ, when the working correlation structure is the identity, Markov, and generalized Markov. Note that, for $\lambda = 3$, the independence model performs well. This is appropriate because, in this case, each $\mathrm{corr}(y_{ij}, y_{ik}) \approx 0$. Table 2 indicates that correctly specifying the generalized Markov structure reduces the mean square error so that, in addition to including an extra parameter that aids in our interpretation of the data, this more general structure also allows for more precise estimation of β.

4.4 *Example*

Here we apply QLS in an analysis of a data set that contains varying numbers of unequally spaced measurements per subject to demonstrate use of the method to select an appropriate correlation structure and to identify potentially important covariates. The data we consider (see Núñez-Anton and Woodworth, 1994, Figure 3) were collected during a study designed to compare different cochlear prostheses implanted in a group of postlingually deafened adults (the Iowa Cochlear Implant (ICI) Project; Gantz et al., 1988). The outcome variable is the percentage of correct responses on a sentence recognition test that was administered at 1, 9, 18, and 30 months postimplantation. Covariates include time of measurement and type of implant (A or B). Due to loss of follow-up, incomplete data were available on treatment groups 0 and 1, comprising 23 and 21 subjects who were implanted with protheses A and B, respectively. To determine if there is a difference in test scores over time between the two treatment groups we used QLS to fit the following model to the full data set:

$$\mathrm{E}(y_{ij}) = \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2 + \beta_3 x_i; \qquad \mathrm{var}(y_{ij}) = \tau; \qquad \mathrm{corr}(Y_i) = R_i(\alpha, \lambda),$$

where y_{ij} is the percentage correct for test j on subject i, x_i is the group indicator variable, t_{ij} is the month in which measurement y_{ij} was made for $j = 1, 2, \ldots, n_i$ and $i = 1, 2, \ldots, 46$. We consider several working correlation structures for $R_i(\alpha, \lambda)$, including the identity, AR(1), Markov, and generalized Markov.

Núñez-Anton and Woodworth (1994) fit the above model to data on subjects who attained at least a 5% improvement over baseline. After confirming multivariate normality for the outcome variable in this subset of the data, they carried out their analysis using the REML approach discussed in Harville (1974). To provide motivation for using an alternative approach, we consider the full data set, which may not be normally distributed, as indicated by an apparent lack of normality in the test scores at 1 and 9 months.

Table 3 contains the regression and correlation parameter estimates for the working correlation structures in Section 3.1. According to the criterion proposed in Section 2.3, the generalized Markov structure is the appropriate structure for these data since it corresponds to the minimum value of the adjusted residual generalized sum of squares $\hat{Q}_a = Q(\hat{\beta}, \hat{\alpha})/(n - p - q)$. We also note that $\hat{Q}_a$ may be used as a rough guide to covariate selection for the final model. For example, if we delete t_{ij}^2 from the model, $\hat{Q}_a = 473.00$ under an assumption of generalized Markov correlation structure, which represents an approximately 8% increase over its value in the original model. This indicates that t_{ij}^2 may be an important covariate to retain in the final model. We also note that fitting the

Table 3
Regression, correlation parameter estimates, and adjusted residual generalized sum of squares for ICI Project data

Working correlation structure	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\alpha}$	$\hat{\lambda}$	$\hat{Q}_a$
Identity	12.97	2.31	−0.05	9.40	—	—	761.23
AR(1)	11.58	2.27	−0.04	11.00	0.64	—	445.43
Markov	11.58	2.32	−0.05	10.73	0.95	—	456.91
Generalized Markov	12.15	2.12	−0.04	11.29	0.84	0.39	439.50

generalized Markov correlation structure allows us to infer, as did Núñez-Anton and Woodworth (1994), that, since $\hat{\lambda} \approx 0$, test scores on one subject tend to stabilize over time. Inclusion of λ in the generalized Markov structure thus aids in our interpretation of the data.

ACKNOWLEDGEMENTS

We would like to thank two anonymous referees whose detailed critiques of an earlier version of this manuscript helped us make substantial changes and improvements. We also thank Drs Nan Laird and John P. Morgan for reviewing the paper. This research was supported in part by U.S. Army research office grant DAAH04-96-1-0070. We also thank Dr Ardythe Morrow for her suggestions and support.

RÉSUMÉ

Les quasi moindre carrés (QLS), une approche statistique marginale via les équations d'estimation généralisées qui sont décrites dans la situation de données équilibrées par Chatangy (1997, *J. Statist. Plann. Inference* **63,** 39–54) permettent l'utilisation d'un grand éventail de structures de corrélation de travail quand on analyse des données présentant une corrélation sérielle. Nous étendons l'application des QLS à des données présentant une corrélation sérielle, espacée de façon inégale, et non équilibrées en utilisant trois modèles utiles de corrélation de travail: l'auto-régressif de premier ordre (AR(1)), le Markov, et la structure de Markov généralisée décrite par Néñez-Anton et Woodworth (1994, *Biometrics* **50,** 445–456). Nous comparons QLS et la formulation originale des équations d'estimation généralisées (GEE) pour ces structures, démontrant que: (i) l'absence de solution aux estimations des paramètres de corrélation peut être un problème; (ii) il est difficile d'obtenir des estimations consistantes des moments des paramètres de corrélation pour la structure de Markov généralisée; (iii) l'utilisation de QLS peut aboutir à une réduction de l'erreur moyenne sur l'estimation des paramètres de régression pour des petits échantillons avec des données modérément corrélées. Pour choisir entre les modèles de corrélation alternatifs, nous proposons un critère qui est basé sur le principe des moindres carrés généralisés. Finalement, des données pour lesquelles la structure de Markov généralisée est appropriée sont analysées, pour démontrer l'utilisation de QLS dans le choix d'une structure de corrélation de travail adaptée et dans l'identification des covariables importantes.

REFERENCES

Chaganty, N. R. (1997). An alternative approach to the analysis of longitudinal data via generalized estimating equations. *Journal of Statistical Planning and Inference* **63,** 39–54.

Dawson, K. S., Gennings, C., and Carter, W. H. (1997). Two graphical techniques useful in detecting correlation structure in repeated measures data. *The American Statistician* **51,** 275–283.

Desmond, A. F. (1997). Optimal estimating functions, quasi-likelihood and statistical modelling. *Journal of Statistical Planning and Inference* **60,** 77–121.

Diggle, P., Liang, K.-Y., and Zeger, S. L. (1994). *Analysis of Longitudinal Data.* Oxford: Oxford University Press.

Fitzmaurice, G. M. (1995). A caveat concerning independence estimating equations with multivariate binary data. *Biometrics* **51,** 309–317.

Fitzmaurice, G. M. and Laird, N. M. (1993). A likelihood-based method for analyzing longitudinal binary responses. *Biometrika* **80,** 141–151.

Gantz, B. J., Tyler, R. S., Knutson, J. F., et al. (1988). Evaluation of five different cochlear implant designs: Audiologic assessment and predictors of performance. *Laryngoscope* **98,** 1100–1106.

Groemping, U. (1994). *GEE: A SAS Macro for Longitudinal Data Analysis,* Version 2.03. Dortmund, Germany: Fachbereich Statistik, Universitaet Dortmund.

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika* **62,** 383–385.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73,** 13–22.

Liang, K. Y., Zeger, S. L., and Qaqish, B. (1992). Multivariate regression analyses for categorical data. *Journal of the Royal Statistical Society, Series B* **54,** 3–40.

Lindsey, J. K. (1993). *Models for Repeated Measurements.* Oxford: Oxford University Press.

Littell, R. C., Milliken, G. A., Stroup, W. W., et al. (1996). *SAS System for Mixed Models.* Cary, North Carolina: SAS Institute.

McDonald, B. W. (1993). Estimating logistic regression parameters for bivariate binary data. *Journal of the Royal Statistical Society, Series B* **54,** 309–317.

Muñoz, A., Carey, V., Schouten, J. P., et al. (1992). A parametric family of correlation structures for the analysis of longitudinal data. *Biometrics* **48,** 733–742.

Núñez-Anton, V. and Woodworth, G. G. (1994). Analysis of longitudinal data with unequally spaced observations and time-dependent correlated errors. *Biometrics* **50,** 445–456.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44,** 1033–1048.

Zhao, L. P. and Prentice, R. L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrics* **77,** 642–684.

Zhao, L. P., Prentice, R. L., and Self, S. G. (1992). Multivariate mean parameter estimation by using a partly exponential model. *Journal of the Royal Statistical Society, Series B* **54,** 805–811.

Received November 1996; *revised September* 1997; *accepted January* 1998.

APPENDIX

The generalized Markov structure $R_i(\alpha, \lambda)$ has a unique Cholesky decomposition given by $\Gamma_i(\alpha,\lambda)\Gamma_i'(\alpha,\lambda)$, where $\Gamma_i(\alpha,\lambda) = [\gamma_{jk}^i]$ is a lower triangular matrix and

$$\gamma_{jk}^i = \begin{cases} 1 & \text{if } j = 1;\ k = 1 \\ \alpha^{e_{i2}+e_{i3}+\cdots+e_{ij}} & \text{if } k = 1;\ j = 2,\ldots,n_i \\ \sqrt{1-\alpha^{2e_{ik}}} & \text{if } k = j;\ j = 2,\ldots,n_i \\ \alpha^{e_{ik+1}+\cdots+e_{ij}}\sqrt{1-\alpha^{2e_{ik}}} & \text{if } k < j;\ j = 2,\ldots,n_i \\ 0 & \text{otherwise.} \end{cases} \tag{5.1}$$

Its inverse, $R_i^{-1}(\alpha,\lambda)$, is a symmetric tridiagonal matrix with unique Cholesky decomposition $L_i(\alpha,\lambda)L_i'(\alpha,\lambda)$, where $L_i(\alpha,\lambda) = [l_{jk}^i]$ and

$$l_{jk}^i = \begin{cases} 1/\sqrt{1-\alpha^{2e_{ij+1}}} & \text{if } k = j;\ j = 1,\ldots,(n_i-1) \\ -\alpha^{e_{ij}}/\sqrt{1-\alpha^{2e_{ij}}} & \text{if } k = j-1;\ j = 2,\ldots,n_i \\ 1 & \text{if } k = j;\ j = n_i; \\ 0 & \text{otherwise.} \end{cases} \tag{5.2}$$

Journal of Applied Statistical Science
Vol. 6, No. 4, pp. 235-243

ISSN 1067-5817

ON INEQUALITIES FOR OUTLIER DETECTION IN STATISTICAL DATA ANALYSIS

N. Rao Chaganty and Akhil K. Vaish

Department of Mathematics and Statistics
Old Dominion University
Norfolk, VA 23529.

SUMMARY

In this paper we obtain some inequalities for quadratic forms involving a symmetric matrix and a positive semidefinite matrix. As special cases of those inequalities, we deduce several known inequalities that are useful for the detection of outliers in statistical data analysis. We also extend Scheffé's S-method of construction of simultaneous confidence intervals for the case where the design matrix is not of full rank and the set of estimable functions are linearly dependent.

Key words and phrases: Inequalities, Positive semidefinite, g-inverse, Outliers, Scheffé's S-method, Multiple comparisons.

AMS 1991 subject classification. Primary: 15A42, 15A18; Secondary: 62J15, 62J10.

1. INTRODUCTION

In a recent article, Olkin (1992) presented an interesting survey of several inequalities that are useful in the detection of outliers in statistical data analysis. In this paper we prove some general inequalities concerning two quadratic forms and deduce most of the inequalities in Olkin (1992) as special cases. As another application to our theorems, we extend the Scheffé's S-method of constructing simultaneous confidence intervals for the case where the design matrix is not of full rank and the set of estimable functions are linearly dependent. The organization of this paper is as follows. In Section 2 we present the main theorems of this paper. Section 3 contains the statistical applications.

2. MAIN RESULTS

We start with the following elementary lemma, stated here without proof since it is well known. It plays an important role in the proofs of the theorems in this paper.

LEMMA 2.1 *Let $\mathbf{C}_{n\times k}$ and $\mathbf{D}_{k\times n}$ be two matrices. Assume that $n \geq k$. Then $(n-k)$ eigenvalues of the matrix $\mathbf{CD}$ are zero and the remaining k eigenvalues of $\mathbf{CD}$, some of which may be zero, coincide with the k eigenvalues of the matrix $\mathbf{DC}$.*

We will now develop some preliminaries before stating the main theorems of this paper. Let $\mathbf{A}$ be a symmetric matrix of order n, $\mathbf{B}$ be a symmetric positive semidefinite matrix of order n and rank equal to k. Let $\mathcal{M}(\mathbf{B})$ denote the column space of $\mathbf{B}$. Let $\mathbf{B} = \mathbf{L}_{n\times k}\mathbf{L}'_{k\times n}$ be the rank factorization of $\mathbf{B}$. Let

$$\begin{aligned} \mathbf{R} &= \mathbf{L}(\mathbf{L}'\mathbf{L})^{-1} \\ \mathbf{A}^* &= \mathbf{R}'\mathbf{A}\mathbf{R} \end{aligned} \tag{2.1}$$

Note that the Moore-Penrose inverse of $\mathbf{B}$ (see Searle (1982), page 220), is given by $\mathbf{B}^+ = \mathbf{R}\mathbf{R}'$. Observe that the column spaces of $\mathbf{B}$, $\mathbf{B}^+$, $\mathbf{R}$ and $\mathbf{L}$ are all equal. Let $\{\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k\}$ be the ordered set of eigenvalues of $\mathbf{A}^*$. Applying Lemma 2.1 for $\mathbf{C} = \mathbf{R}$ and $\mathbf{D} = \mathbf{R}'\mathbf{A}$ we can see that the set of eigenvalues of the matrix $\mathbf{B}^+\mathbf{A}$ is given by $\{\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k, 0, \ldots, 0\}$. It is possible that some of the $\lambda_i's$ may equal zero and also, all the λ_i's may be negative. Therefore λ_1 need not be the largest eigenvalue of $\mathbf{B}^+\mathbf{A}$. Similarly, λ_k need not be the smallest eigenvalue of $\mathbf{B}^+\mathbf{A}$. In fact the largest eigenvalue of $\mathbf{B}^+\mathbf{A}$ is given by $\max\{0, \lambda_1\}$ and the smallest eigenvalue of $\mathbf{B}^+\mathbf{A}$ is equal to $\min\{0, \lambda_k\}$. We are now ready to state an inequality concerning two quadratic forms.

THEOREM 2.2 *Let $\mathbf{A}$ be a symmetric matrix of order n. Let $\mathbf{B}$ be a symmetric positive semidefinite matrix of order n and rank equal to k. Let $\{\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k, 0, \ldots, 0\}$ be the set of n eigenvalues of $\mathbf{B}^+\mathbf{A}$. Then*

$$\lambda_k \;\; \mathbf{y}'\mathbf{B}\,\mathbf{y} \leq \mathbf{y}'\mathbf{A}\,\mathbf{y} \leq \lambda_1 \;\; \mathbf{y}'\mathbf{B}\,\mathbf{y} \tag{2.2}$$

for all $\mathbf{y} \in \mathcal{M}(\mathbf{B})$. There exists an eigenvector $\mathbf{y}_i$ of $\mathbf{B}^+\mathbf{A}$ corresponding to the eigenvalue λ_i such that $\mathbf{y}_i \in \mathcal{M}(\mathbf{B})$ for $1 \leq i \leq k$. Further, equality holds in the first inequality and in the second inequality of (2.2) if we choose $\mathbf{y}$ to be equal to $\mathbf{y}_k$ and $\mathbf{y}_1$ respectively.

Proof. Let $\mathbf{B} = \mathbf{L}\mathbf{L}'$ be the rank factorization of $\mathbf{B}$. Let $\mathbf{R}$ and $\mathbf{A}^*$ be as defined in (2.1). Since λ_1 and λ_k are the largest and the smallest eigenvalues of $\mathbf{A}^*$, by a well known inequality (see (1f.2.1) of Rao (1973), page 62) we have

$$\lambda_k\,\mathbf{v}'\mathbf{v} \leq \mathbf{v}'\mathbf{A}^*\mathbf{v} \leq \lambda_1 \;\; \mathbf{v}'\mathbf{v} \tag{2.3}$$

for all $\mathbf{v} \in \Re^k$. Let $\mathbf{y}$ be a vector in $\mathcal{M}(\mathbf{B})$ and let $\mathbf{v} = \mathbf{L}'\mathbf{y}$. It is easy to verify that $\mathbf{y} = \mathbf{R}\mathbf{v}$, since $\mathbf{y}$ is also in the column space of $\mathbf{L}$. Thus we have

$$\begin{aligned} \mathbf{v}'\mathbf{v} &= \mathbf{y}'\mathbf{B}\mathbf{y} \\ \mathbf{v}'\mathbf{A}^*\mathbf{v} &= \mathbf{y}'\mathbf{A}\mathbf{y}. \end{aligned} \tag{2.4}$$

The assertion (2.2) now follows from (2.3) and (2.4). We now proceed to show that the two inequalities in (2.2), become equalities for appropriate choices of $\mathbf{y}$. For $1 \leq i \leq k$, let $\mathbf{v}_i \neq 0$ be an eigenvector of $\mathbf{A}^*$ corresponding to the eigenvalue λ_i and let $\mathbf{y}_i = \mathbf{R}\mathbf{v}_i$. Note that $\mathbf{y}_i \neq 0$, since $\mathbf{R}$ is of full column rank and $\mathbf{v}_i \neq 0$. We also have

$$\begin{aligned} \mathbf{B}^+\mathbf{A}\mathbf{y}_i &= \mathbf{R}\mathbf{R}'\mathbf{A}\mathbf{R}\mathbf{v}_i \\ &= \mathbf{R}\mathbf{A}^*\mathbf{v}_i = \lambda_i\mathbf{R}\mathbf{v}_i \end{aligned}$$

$$= \lambda_i \mathbf{y}_i. \tag{2.5}$$

Thus $\mathbf{y}_i$ is an eigenvector of $\mathbf{B}^+\mathbf{A}$, corresponding to the eigenvalue λ_i. Clearly $\mathbf{y}_i \in \mathcal{M}(\mathbf{B})$ since it is in the column space of $\mathbf{R}$. Therefore from (2.4) we have $\mathbf{v}_i'\mathbf{A}^*\mathbf{v}_i = \mathbf{y}_i'\mathbf{A}\mathbf{y}_i$ and $\mathbf{v}_i'\mathbf{v}_i = \mathbf{y}_i'\mathbf{B}\mathbf{y}_i$. Thus $\mathbf{y}_i'\mathbf{A}\mathbf{y}_i = \lambda_i\mathbf{y}_i'\mathbf{B}\mathbf{y}_i$ for all $1 \le i \le k$. Therefore the first and second inequalities in (2.2) become equalities if we choose $\mathbf{y}$ to be equal to $\mathbf{y}_k$ and $\mathbf{y}_1$ respectively. This completes the proof of Theorem 2.2. □

In the case where $\lambda_1 = \cdots = \lambda_k$, from (2.2) we have $\mathbf{y}'\mathbf{A}\mathbf{y} = \lambda_1\mathbf{y}'\mathbf{B}\mathbf{y}$ for all $\mathbf{y} \in \mathcal{M}(\mathbf{B})$. The following example shows that this need not be true for vectors $\mathbf{y}$ which are not in $\mathcal{M}(\mathbf{B})$. The example also shows that (2.2) need not be true for vectors $\mathbf{y}$ not in $\mathcal{M}(\mathbf{B})$.

EXAMPLE 2.3 Let $\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ and $\mathbf{B} = \begin{pmatrix} 1/4 & 1/4 \\ 1/4 & 1/4 \end{pmatrix}$. Clearly, $\mathbf{B}$ is positive semidefinite matrix of rank $k = 1$. The Moore-Penrose inverse of $\mathbf{B}$ is given by $\mathbf{B}^+ = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$. It is easy to verify that the set of eigenvalues of $\mathbf{B}^+\mathbf{A}$ is given by $\{1, 0\}$ and therefore λ_1 equals 1. Consider the vector $\mathbf{y}' = (2, 0)$ which is not in the column space of $\mathbf{B}$. A little calculation shows that $\mathbf{y}'\mathbf{B}\mathbf{y} = 1$ and $\mathbf{y}'\mathbf{A}\mathbf{y} = 4$ and hence $\mathbf{y}'\mathbf{A}\mathbf{y} > \lambda_1\mathbf{y}'\mathbf{B}\mathbf{y}$. Similarly, for $\mathbf{y}' = (0, 2)$ we have $\mathbf{y}'\mathbf{A}\mathbf{y} < \lambda_1\mathbf{y}'\mathbf{B}\mathbf{y}$. Therefore this example shows that the inequalities in (2.2) need not hold for all $\mathbf{y}$.

The next theorem gives sufficient condition for the inequality (2.2) to hold for all $\mathbf{y} \in \Re^n$.

THEOREM 2.4 *Let $\mathbf{A}$ be a symmetric matrix of order n. Let $\mathbf{B}$ be a symmetric positive semidefinite matrix of order n and rank equal to k. Let $\{\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_k, 0, \ldots, 0\}$ be the set of n eigenvalues of $\mathbf{B}^+\mathbf{A}$. If $\mathcal{M}(\mathbf{A}) \subseteq \mathcal{M}(\mathbf{B})$ then the matrices $\lambda_1\mathbf{B} - \mathbf{A}$ and $\mathbf{A} - \lambda_k\mathbf{B}$ are positive semidefinite.*

Proof. Fix $\mathbf{y} \in \Re^n$. Then we can write $\mathbf{y} = \mathbf{y}_b + \mathbf{y}_b^\perp$, where $\mathbf{y}_b$ is the projection of $\mathbf{y}$ onto the column space of $\mathbf{B}$ and $\mathbf{y}_b^\perp = \mathbf{y} - \mathbf{y}_b$. Note that $\mathbf{B}\mathbf{y}_b^\perp = 0$. If $\mathcal{M}(\mathbf{A}) \subseteq \mathcal{M}(\mathbf{B})$ we also have $\mathbf{A}\mathbf{y}_b^\perp = 0$. Therefore,

$$\begin{aligned} \mathbf{y}'\mathbf{A}\mathbf{y} &= \mathbf{y}_b'\mathbf{A}\mathbf{y}_b \\ \mathbf{y}'\mathbf{B}\mathbf{y} &= \mathbf{y}_b'\mathbf{B}\mathbf{y}_b. \end{aligned} \tag{2.6}$$

Since $\mathbf{y}_b \in \mathcal{M}(\mathbf{B})$ by (2.2) of Theorem 2.2 we have

$$\lambda_k \ \mathbf{y}_b'\mathbf{B}\mathbf{y}_b \le \mathbf{y}_b'\mathbf{A}\mathbf{y}_b \le \lambda_1 \ \mathbf{y}_b'\mathbf{B}\mathbf{y}_b. \tag{2.7}$$

Combining (2.6) and (2.7) we get

$$\lambda_k \ \mathbf{y}'\mathbf{B}\mathbf{y} \le \mathbf{y}'\mathbf{A}\mathbf{y} \le \lambda_1 \ \mathbf{y}'\mathbf{B}\mathbf{y}. \tag{2.8}$$

Since $\mathbf{y} \in \Re^n$ is arbitrary, (2.8) shows that the matrices $\lambda_1\mathbf{B} - \mathbf{A}$ and $\mathbf{A} - \lambda_k\mathbf{B}$ are positive semidefinite. □

The next lemma shows that for any two symmetric matrices $\mathbf{A}$ and $\mathbf{B}$, if $\mathcal{M}(\mathbf{A}) \subseteq \mathcal{M}(\mathbf{B})$ then the set of eigenvalues of $\mathbf{B}^-\mathbf{A}$ is invariant of the choice of the g-inverse $\mathbf{B}^-$ of $\mathbf{B}$. Thus we can replace $\mathbf{B}^+$ by any g-inverse $\mathbf{B}^-$ of $\mathbf{B}$ in the statement of Theorem 2.4.

LEMMA 2.5 *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be two symmetric matrices both of order* n. *If* $\mathcal{M}(\mathbf{A}) \subseteq \mathcal{M}(\mathbf{B})$ *then the set of eigenvalues of* $\mathbf{B}^- \mathbf{A}$ *is invariant of the choice of the g-inverse,* $\mathbf{B}^-$, *of* $\mathbf{B}$.

Proof. Let $\mathbf{A}$ and $\mathbf{B}$ be two symmetric matrices of order n such that $\mathcal{M}(\mathbf{A}) \subseteq \mathcal{M}(\mathbf{B})$. By spectral decomposition, there exists orthogonal matrix $\mathbf{P}$ such that

$$\mathbf{A} = \mathbf{P} \begin{pmatrix} \mathbf{\Lambda} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix} \mathbf{P}' = \mathbf{P}_1 \mathbf{\Lambda} \mathbf{P}_1'$$

where $\mathbf{\Lambda}$ is a diagonal matrix, $\mathbf{O}$ is a null matrix and $\mathbf{P} = [\mathbf{P}_1 \ \mathbf{P}_2]$ is the partition of $\mathbf{P}$, depending on the rank of the matrix $\mathbf{A}$. Since $\mathcal{M}(\mathbf{A}) = \mathcal{M}(\mathbf{P}_1)$ and $\mathcal{M}(\mathbf{A}) \subseteq \mathcal{M}(\mathbf{B})$, we have $\mathcal{M}(\mathbf{P}_1) \subseteq \mathcal{M}(\mathbf{B})$. Hence we can write $\mathbf{P}_1 = \mathbf{B}\,\mathbf{U}$ for some matrix $\mathbf{U}$. Therefore,

$$\mathbf{A} = \mathbf{P}_1 \mathbf{\Lambda} \mathbf{P}_1' = \mathbf{B}\,\mathbf{U}\,\mathbf{\Lambda}\,\mathbf{U}'\,\mathbf{B} = \mathbf{B}\,\mathbf{V}\,\mathbf{B} \tag{2.9}$$

where $\mathbf{V} = \mathbf{U}\,\mathbf{\Lambda}\,\mathbf{U}'$, is a symmetric matrix. Let $\mathbf{B}^-$ be a g-inverse of $\mathbf{B}$. If we choose $\mathbf{C} = \mathbf{B}^-\,\mathbf{B}\,\mathbf{V}$ and $\mathbf{D} = \mathbf{B}$, then by Lemma 2.1 we have that the set of eigenvalues of $\mathbf{B}^-\mathbf{A}$ is exactly same as the set of eigenvalues of the matrix $\mathbf{B}\,\mathbf{V}$. Thus, the eigenvalues of $\mathbf{B}^-\,\mathbf{A}$ do not depend on the choice of the g-inverse $\mathbf{B}^-$, of $\mathbf{B}$. This completes the proof of the lemma. □

The following example shows that the conclusion of Lemma 2.5 need not be true if we do not assume that $\mathcal{M}(\mathbf{A})$ is contained in $\mathcal{M}(\mathbf{B})$.

EXAMPLE 2.6 Consider the matrices $\mathbf{A}$ and $\mathbf{B}$ as in Example 2.3. It is easy to verify that $\mathcal{M}(\mathbf{A})$ is not contained in $\mathcal{M}(\mathbf{B})$. We have seen that in Example 2.3 the set of eigenvalues of $\mathbf{B}^+\mathbf{A}$ is given by $\{1, 0\}$. Consider another g-inverse $\mathbf{B}^- = \begin{pmatrix} 2 & 0 \\ 2 & 0 \end{pmatrix}$, of $\mathbf{B}$. We can easily verify that the set of eigenvalues of $\mathbf{B}^-\mathbf{A}$ is given by $\{2, 0\}$, which is different from the set of eigenvalues of $\mathbf{B}^+\mathbf{A}$. Thus, the conclusion of Lemma 2.5 need not be true if $\mathcal{M}(\mathbf{A})$ is not contained in $\mathcal{M}(\mathbf{B})$.

Let $\mathbf{b}$ be a vector in $\Re^n$. Theorems 2.2 and 2.4 restricted to the matrix $\mathbf{A} = \mathbf{b}\mathbf{b}'$ give rise to several interesting inequalities. We treat this special case in Theorem 2.7 below. In Section 3 we will use Theorem 2.7 to prove several inequalities that are useful in the detection of outliers in statistical data analysis.

THEOREM 2.7 *Let* $\mathbf{B}$ *be a symmetric positive semidefinite matrix of order* n. *Let* $\mathbf{B}^+$ *be the Moore-Penrose inverse of* $\mathbf{B}$. *Let* $\mathcal{M}(\mathbf{B})$ *denote the column space of* $\mathbf{B}$. *If* $\mathbf{b}$ *is an* $n \times 1$ *vector then*

$$(\mathbf{b}'\mathbf{y})^2 \leq \mathbf{b}'\mathbf{B}^+\mathbf{b} \quad \mathbf{y}'\mathbf{B}\mathbf{y} \tag{2.10}$$

for all $\mathbf{y} \in \mathcal{M}(\mathbf{B})$. *Moreover, equality holds in (2.10) if we choose* $\mathbf{y} = \mathbf{B}^+\mathbf{b}$. *Also, if rank of* $\mathbf{B}$ *equals 1, then equality holds in (2.10) for all* $\mathbf{y} \in \mathcal{M}(\mathbf{B})$. *If* $\mathbf{b} \in \mathcal{M}(\mathbf{B})$ *then (2.10) holds for all* $\mathbf{y} \in \Re^n$, *equivalently, the matrix* $(\mathbf{b}'\mathbf{B}^+\mathbf{b})\ \mathbf{B} - \mathbf{b}\mathbf{b}'$ *is positive semidefinite.*

Proof. Let $\mathbf{b}$ be an $n \times 1$ vector. Let us choose $\mathbf{A} = \mathbf{b}\,\mathbf{b}'$ in Theorems 2.2 and 2.4. Letting $\mathbf{C} = \mathbf{B}^+\mathbf{b}$ and $\mathbf{D} = \mathbf{b}'$ in Lemma 2.1 we can see that the set of eigenvalues of $\mathbf{B}^+\mathbf{A}$ is given by $\{\mathbf{b}'\mathbf{B}^+\mathbf{b}, 0, \ldots, 0\}$. Let the rank of $\mathbf{B}$ be equal to k. Then, in the notation of Theorem 2.2, the eigenvalue λ_1 equals $\mathbf{b}'\mathbf{B}^+\mathbf{b}$ and $\lambda_k = \lambda_1$ if $k = 1$ and $\lambda_k = 0$ if $k \geq 2$. Therefore Theorem 2.7 follows from Theorems 2.2 and 2.4. □

The following Corollary 2.8 is an easy consequence of Theorem 2.7. We will apply this corollary in Section 3, to extend Scheffé's S-method of constructing simultaneous confidence intervals, when the design matrix is not of full rank and the set of estimable functions are linearly dependent.

COROLLARY 2.8 *Let* $\mathbf{B}$ *be a symmetric positive semidefinite matrix of order* n *and* $\mathbf{B}^-$ *be a g-inverse of* $\mathbf{B}$. *Let* $\mathbf{b} \in \mathcal{M}(\mathbf{B})$ *then* $\eta\mathbf{B} - \mathbf{b}\mathbf{b}'$ *is positive semidefinite if and only if* $\eta \geq \mathbf{b}'\mathbf{B}^-\mathbf{b}$.

Proof. Let $\mathbf{b} \in \mathcal{M}(\mathbf{B})$. It is easy to verify that $\mathbf{b}'\mathbf{B}^-\mathbf{b} = \mathbf{b}'\mathbf{B}^+\mathbf{b}$ for any choice of the g-inverse $\mathbf{B}^-$, of $\mathbf{B}$. Suppose that $\eta \geq \mathbf{b}'\mathbf{B}^+\mathbf{b}$, then from Theorem 2.7 we have

$$\eta\mathbf{y}'\mathbf{B}\mathbf{y} \geq \mathbf{b}'\mathbf{B}^+\mathbf{b}\ \ \mathbf{y}'\mathbf{B}\mathbf{y} \geq \mathbf{y}'\mathbf{b}\mathbf{b}'\mathbf{y} \tag{2.11}$$

for all $\mathbf{y} \in \Re^n$. Therefore $\eta\mathbf{B} - \mathbf{b}\mathbf{b}'$ is positive semidefinite. The other implication follows easily, if we choose $\mathbf{y} = \mathbf{B}^+\mathbf{b}$. □

Theorem 2.7 is essentially asserting that if $\mathbf{B}$ is a positive semidefinite matrix and $\mathbf{b} \in \Re^n$ then

$$\sup_{\substack{\mathbf{y} \in \mathcal{M}(\mathbf{B}) \\ \mathbf{y} \neq 0}} \frac{(\mathbf{b}'\mathbf{y})^2}{\mathbf{y}'\mathbf{B}\mathbf{y}} = \mathbf{b}'\mathbf{B}^+\mathbf{b}. \tag{2.12}$$

This generalizes the result contained in (11), Appendix A4 of Seber (1977), where the above equality (2.12) was obtained for positive definite matrix $\mathbf{B}$. Note that if $\mathbf{A}$ is a symmetric matrix and $\mathbf{B}$ is a positive semidefinite matrix then the conclusion of Theorem 2.2 can be restated as

$$\sup_{\substack{\mathbf{y} \in \mathcal{M}(\mathbf{B}) \\ \mathbf{y} \neq 0}} \frac{\mathbf{y}'\mathbf{A}\mathbf{y}}{\mathbf{y}'\mathbf{B}\mathbf{y}} = \lambda_1 \qquad \inf_{\substack{\mathbf{y} \in \mathcal{M}(\mathbf{B}) \\ \mathbf{y} \neq 0}} \frac{\mathbf{y}'\mathbf{A}\mathbf{y}}{\mathbf{y}'\mathbf{B}\mathbf{y}} = \lambda_k \tag{2.13}$$

where λ_1 and λ_k are the eigenvalues of $\mathbf{B}^+\mathbf{A}$ as defined in Theorem 2.2. A similar representation is also true for the other eigenvalues λ_p, $2 \leq p \leq (k-1)$ and is given by the following theorem.

THEOREM 2.9 *Let* $\mathbf{A}$ *be a symmetric matrix of order* n. *Let* $\mathbf{B}$ *be symmetric positive semidefinite matrix of order* n *and rank equal to* k. *Let* $\{\lambda_1 \geq \cdots \geq \lambda_k, 0, \ldots, 0\}$ *be the set of* n *eigenvalues of* $\mathbf{B}^+\mathbf{A}$. *Then there exist eigenvectors* $\{\mathbf{y}_1, \ldots, \mathbf{y}_k\}$ *of* $\mathbf{B}^+\mathbf{A}$, *corresponding to the eigenvalues* $\{\lambda_1, \ldots, \lambda_k\}$ *such that* $\mathbf{y}_i \in \mathcal{M}(\mathbf{B})$, $\mathbf{y}_i'\mathbf{B}\mathbf{y}_j = 0$, $1 \leq i \neq j \leq k$. *Further*

$$\sup_{\{\mathbf{y} \in \mathcal{B}_p,\ \mathbf{y} \neq 0\}} \frac{\mathbf{y}'\mathbf{A}\mathbf{y}}{\mathbf{y}'\mathbf{B}\mathbf{y}} = \lambda_p \tag{2.14}$$

where $\mathcal{B}_p = \{\mathbf{y} \in \mathcal{M}(\mathbf{B}) : \mathbf{y}_i'\mathbf{B}\mathbf{y} = 0,\ 1 \leq i \leq (p-1)\}$, *for* $2 \leq p \leq (k-1)$.

Proof. Let $\mathbf{A}$ be a symmetric matrix and $\mathbf{B}$ be a symmetric positive semidefinite matrix of rank equal to k and $\mathbf{B}^+$ denote the Moore-Penrose inverse of $\mathbf{B}$. Let $\mathbf{L}$ and $\mathbf{R}$ and $\mathbf{A}^*$ be as defined in (2.1). Let $\{\lambda_1 \geq \cdots \geq \lambda_k\}$ be the set of ordered eigenvalues of $\mathbf{A}^*$ and $\mathbf{v}_1, \ldots, \mathbf{v}_k$ be corresponding orthogonal eigenvectors. By Theorem 1 of Bellman (1970), page 113, we have

$$\sup_{\substack{\{\mathbf{v} \in \Re^k : \mathbf{v}_i'\mathbf{v} = 0\} \\ \mathbf{v} \neq 0 \qquad 1 \leq i \leq (p-1)}} \frac{\mathbf{v}'\mathbf{A}^*\mathbf{v}}{\mathbf{v}'\mathbf{v}} = \lambda_p \qquad \text{for } 2 \leq p \leq (k-1). \tag{2.15}$$

Let $\mathbf{y} = \mathbf{R}\mathbf{v}$, then as $\mathbf{v}$ varies in $\Re^k$, the vector $\mathbf{y}$ varies in $\mathcal{M}(\mathbf{B})$ and by (2.4) we have $\mathbf{v}'\mathbf{v} = \mathbf{y}'\mathbf{B}\mathbf{y}$ and $\mathbf{v}'\mathbf{A}^*\mathbf{v} = \mathbf{y}'\mathbf{A}\mathbf{y}$. Let us define $\mathbf{y}_i = \mathbf{R}\mathbf{v}_i$ for $1 \leq i \leq k$. Then by Theorem 2.2, $\mathbf{y}_i$ is the eigenvector of $\mathbf{B}^+\mathbf{A}$ corresponding to the eigenvalue λ_i. Further $\mathbf{y}_i'\mathbf{B}\mathbf{y}_j = 0$, since $\mathbf{v}_i = \mathbf{L}'\mathbf{y}_i$ and $\mathbf{v}_i'\mathbf{v}_j = 0$ for $1 \leq i \neq j \leq k$. The identity (2.14) now follows from (2.15). □

3. STATISTICAL APPLICATIONS

In this section we present some applications of the theorems in Section 2. Our first application deals with some inequalities that are useful for the detection of outliers in statistical data. As a second application we extend Scheffé's S-method of constructing simultaneous confidence intervals, when the design matrix is not of full rank and the set of given estimable functions are linearly dependent.

APPLICATION 3.1 In a recent paper Olkin (1992) considered the following problem, that is of interest in the detection of outliers. Given the mean and standard deviation of a finite sample, find the maximum deviation of any particular observation from the sample mean as a multiple of the sample standard deviation. More specifically, let $\{y_1, \ldots, y_n\}$ be a sample of n observations. The problem is to find the minimum value of c such that

$$(y_k - \bar{y})^2 \leq c \sum_{i=1}^{n} (y_i - \bar{y})^2, \quad k = 1, \ldots, n, \tag{3.1}$$

where $\bar{y} = \sum_{i=1}^{n} y_i/n$, is the sample mean. The above problem and its solution that the best value of c equals $(n-1)/n$ was first brought into the limelight of statistics by Samuelson (1968). The inequality (3.1) with $c = (n-1)/n$ is now popularly known as Samuelson's inequality. Olkin (1992) gave an interesting survey of the known proofs of Samuelson's inequality and raised the question whether there is room for yet another proof. He then gave a new proof with some generalizations. We now show that Samuelson's inequality and several other inequalities in Olkin (1992) follow from our theorems of Section 2.

Let $\mathbf{e}' = (1, \ldots, 1)$ and $\mathbf{B} = \mathbf{I}_n - \frac{1}{n}\mathbf{e}\mathbf{e}'$, where $\mathbf{I}_n$ is the identity matrix. Note that $\mathbf{B}$ is symmetric, idempotent matrix. Hence $\mathbf{B}$ is positive semidefinite and $\mathbf{B}^+ = \mathbf{B}$. Fix $1 \leq k \leq n$. Consider the vector $\mathbf{b}_1$, where the jth component is given by

$$b_{1j} = \begin{cases} 1 - (1/n) & \text{if } j = k \\ -1/n & \text{if } j \neq k \end{cases} \tag{3.2}$$

Since $\mathbf{b}_1'\mathbf{e} = 0$ we have $\mathbf{b}_1 \in \mathcal{M}(\mathbf{B})$. Also, $\mathbf{b}_1'\mathbf{B}^+\mathbf{b}_1 = \mathbf{b}_1'\mathbf{b}_1 = (n-1)/n$. If we choose $\mathbf{b} = \mathbf{b}_1$, by the last assertion of Theorem 2.7 we have $((n-1)/n)\mathbf{B} - \mathbf{b}\mathbf{b}'$ is positive semidefinite. Thus for any $\mathbf{y} \in \Re^n$ we get

$$\frac{(n-1)}{n}\mathbf{y}'\mathbf{B}\mathbf{y} \geq \mathbf{y}'\mathbf{b}_1\mathbf{b}_1'\mathbf{y} \tag{3.3}$$

which is equivalent to Samuleson's inequality:

$$(y_k - \bar{y})^2 \leq \frac{(n-1)}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2. \tag{3.4}$$

In a similar fashion we can deduce inequalities (2.3) and (2.4) of Olkin (1992) as a consequence of Theorem 2.7 if we choose $\mathbf{b} = \mathbf{b}_2$, and $\mathbf{b} = \mathbf{b}_3$ respectively, where the jth component of the vectors $\mathbf{b}_2$ and $\mathbf{b}_3$ are given by

$$b_{2j} = \begin{cases} (1/k) - (1/n) & \text{if } 1 \leq j \leq k \\ -1/n & \text{if } k+1 \leq j \leq n \end{cases} \tag{3.5}$$

$$b_{3j} = \begin{cases} 1/k & \text{if } 1 \le j \le k \\ -1/r & \text{if } k+1 \le j \le k+r \\ 0 & \text{if } k+r < j \le n. \end{cases} \tag{3.6}$$

Also the inequality in Olkin (1992) involving Gini mean difference, due to Nair (1956), follows from Theorem 2.7 if we choose $\mathbf{b} = \mathbf{b}_4$, where the jth component of $\mathbf{b}_4$ is given by

$$b_{4j} = \frac{2(2j-n-1)}{n(n-1)} \quad \text{for } 1 \le j \le n. \tag{3.7}$$

Let us choose the vector $\mathbf{b} = \mathbf{b}_5$ in Theorem 2.7, where the jth component of $\mathbf{b}_5$ is given by

$$b_{5j} = \begin{cases} -1 & \text{if } j = 1 \\ 1 & \text{if } j = n \\ 0 & \text{otherwise.} \end{cases} \tag{3.8}$$

Clearly $\mathbf{b}_5 \in \mathcal{M}(\mathbf{B})$ and $\mathbf{b}_5' \mathbf{B}^+ \mathbf{b}_5 = \mathbf{b}_5' \mathbf{b}_5 = 2$. Thus from Theorem 2.7 we have $2\mathbf{B} - \mathbf{b}_5 \mathbf{b}_5'$ is positive semidefinite. For a vector $\mathbf{y}' = (y_1, \ldots, y_n)$, let $\tilde{\mathbf{y}}' = (y_{(1)}, \ldots, y_{(n)})$ where $y_{(i)}$'s are the ordered values of the components of $\mathbf{y}$. Since $2\mathbf{B} - \mathbf{b}_5 \mathbf{b}_5'$ is positive semidefinite we have

$$2\tilde{\mathbf{y}}' \mathbf{B} \tilde{\mathbf{y}} \ge \tilde{\mathbf{y}}' \mathbf{b}_5 \mathbf{b}_5 \tilde{\mathbf{y}} \tag{3.9}$$

which after simplification reduces to an inequality, due to Thompson (1955), given by

$$(y_{(n)} - y_{(1)})^2 \le 2 \sum_{i=1}^{n} (y_i - \bar{y})^2. \tag{3.10}$$

We will now show that the multidimensional inequalities contained in Olkin (1992) can also be deduced from our theorems. Let $\mathbf{W}$ be a matrix of order $l \times n$ such that $\mathbf{W}\mathbf{e} = 0$ and $\mathbf{W}\mathbf{W}' = \mathbf{I}_l$. Let $\mathbf{B} = \mathbf{I}_n - \frac{1}{n}\mathbf{e}\mathbf{e}'$ be as before and let $\mathbf{A} = \mathbf{W}'\mathbf{W}$. Since $\mathbf{B}^+ = \mathbf{B}$ and $\mathbf{A}\mathbf{e} = 0$ we have $\mathcal{M}(\mathbf{A}) \subseteq \mathcal{M}(\mathbf{B})$ and $\mathbf{B}^+\mathbf{A} = \mathbf{A}$. Hence the largest eigenvalue of $\mathbf{B}^+\mathbf{A}$ equals the largest eigenvalue of $\mathbf{A}$ and which in turn equals the largest eigenvalue of $\mathbf{W}\mathbf{W}'$. Therefore $\lambda_1 = 1$ for this choice of $\mathbf{B}$ and $\mathbf{A}$. Therefore by Theorem 2.4, we have $\mathbf{B} - \mathbf{A}$ is positive semidefinite. Thus we get

$$\mathbf{y}'\mathbf{W}'\mathbf{W}\mathbf{y} \le \mathbf{y}'\left(\mathbf{I}_n - \frac{1}{n}\mathbf{e}\mathbf{e}'\right)\mathbf{y} \quad \textit{for all } \mathbf{y} \in R^n. \tag{3.11}$$

Thus for any $m \times n$ matrix $\mathbf{Z}$, the matrix

$$\mathbf{Z}\left(\mathbf{I}_n - \frac{1}{n}\mathbf{e}\mathbf{e}'\right)\mathbf{Z}' - \mathbf{Z}\mathbf{W}'\mathbf{W}\mathbf{Z}' \tag{3.12}$$

is positive semidefinite. Hence inequality (3.6) in Olkin (1992) holds.

APPLICATION 3.2 Our second application deals with multiple comparison procedures in linear models. One of the most important problems in multiple comparisons is the problem of construction of simultaneous confidence intervals for a given set of estimable functions. Among the several methods available, Scheffé's technique has been the most popular and widely used method for the construction of simultaneous confidence intervals. A very nice description of Scheffé's S-method can be found in Seber (1977), page 128. In most texts the S-method is usually described assuming that the design matrix is of full rank and the set of given estimable functions are linearly independent. However,

this is rarely the case in practice. As an important application of the results of Section 2 we now show that Scheffé's S-method can be extended to the case where the design matrix is not of full rank and the set of estimable functions are linearly dependent.

Consider the linear model $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, where $\mathbf{y}$ is an $n \times 1$ vector of observations, β is a $p \times 1$ vector of parameters, $\mathbf{X}$ is a design matrix of order $n \times p$ and ε is a $n \times 1$ vector of random errors. Let us assume that ε is distributed as multivariate normal with mean $\mathbf{0}$ and variance-covariance matrix $\sigma^2\mathbf{I}_n$. Assume that the rank of $\mathbf{X}$ is r, where $r < p$. Consider s estimable functions $\mathbf{K}'\beta$, where $\mathbf{K}_{p\times s}$ is a matrix of rank $q < s$. It is well known that the condition of estimability is equivalent to $\mathcal{M}(\mathbf{K}) \subseteq \mathcal{M}(\mathbf{X}'\mathbf{X})$. Let $\mathbf{G}$ be a g-inverse of $\mathbf{X}'\mathbf{X}$ and $(n-r)\hat{\sigma}^2 = \mathbf{y}'(\mathbf{I}_n - \mathbf{X}\mathbf{G}\mathbf{X}')\mathbf{y}$. From Theorem 4.6 of Seber (1977) it follows that the statistic

$$F = \frac{(\mathbf{K}'\hat{\beta} - \mathbf{K}'\beta)'\,(\mathbf{K}'\,\mathbf{G}\,\mathbf{K})^{-}\,(\mathbf{K}'\hat{\beta} - \mathbf{K}'\beta)/q}{\hat{\sigma}^2} \tag{3.13}$$

has an F-distribution with q and $(n-r)$ degrees of freedom, where $\hat{\beta}$ is any solution to the equation $\mathbf{X}'\mathbf{X}\beta = \mathbf{X}'\mathbf{y}$. Let $F^{\alpha}_{q,n-r}$ be the $100(1-\alpha)$ percentile of the F-distribution with q and $(n-r)$ degrees of freedom, then from (3.13) we have

$$\begin{aligned} 1-\alpha &= Pr\left(F \le F^{\alpha}_{q,n-r}\right) \\ &= Pr\left((\mathbf{K}'\hat{\beta} - \mathbf{K}'\beta)'\,(\mathbf{K}'\mathbf{G}\mathbf{K})^{-}\,(\mathbf{K}'\hat{\beta} - \mathbf{K}'\beta) \le q\,\hat{\sigma}^2\,F^{\alpha}_{q,n-r}\right) \\ &= Pr\left(\ \mathbf{b}'\mathbf{B}^{-}\mathbf{b}\ \le\ \eta\ \right) \end{aligned} \tag{3.14}$$

where $\eta = q\hat{\sigma}^2\,F^{\alpha}_{q,n-r}$, $\mathbf{B} = \mathbf{K}'\,\mathbf{G}\,\mathbf{K}$ and $\mathbf{b} = \mathbf{K}'(\hat{\beta} - \beta)$. Note that $\mathbf{b} \in \mathcal{M}(\mathbf{B})$ since $\mathbf{b} \in \mathcal{M}(\mathbf{K}')$ and from Lemma 3.3 below we have $\mathcal{M}(\mathbf{K}') = \mathcal{M}(\mathbf{K}'\,\mathbf{G}\,\mathbf{K})$. Since $\mathbf{B}$ is a symmetric positive semidefinite matrix and $\mathbf{b} \in \mathcal{M}(\mathbf{B})$, by Corollary 2.8 we have that (3.14) is equivalent to

$$\begin{aligned} 1-\alpha &= Pr\left(\ \mathbf{h}'\mathbf{b}\,\mathbf{b}'\mathbf{h}\ \le\ \eta\mathbf{h}'\mathbf{B}\,\mathbf{h}\quad \text{for all } \mathbf{h}\ \right) \\ &= Pr\left(\ |\mathbf{h}'(\mathbf{K}'\hat{\beta} - \mathbf{K}'\beta)|\ \le\ \sqrt{\eta\mathbf{h}'\mathbf{B}\,\mathbf{h}}\quad \text{for all } \mathbf{h}\right). \end{aligned} \tag{3.15}$$

We therefore have a simultaneous confidence intervals for any linear function $\mathbf{h}'(\mathbf{K}'\beta)$ of the estimable functions $\mathbf{K}'\beta$, namely,

$$\mathbf{h}'(\mathbf{K}'\hat{\beta}) \pm (q\,F^{\alpha}_{q,n-r})^{1/2}\,\hat{\sigma}\,\sqrt{\mathbf{h}'(\mathbf{K}'\mathbf{G}\mathbf{K})\mathbf{h}} \tag{3.16}$$

such that the overall probability for the whole class of such intervals is equal to $(1-\alpha)$.

We have used the following lemma in Application 3.2.

LEMMA 3.3 *Let* $\mathbf{K}$ *and* $\mathbf{X}$ *be as in Application 3.2. Suppose that* $\mathcal{M}(\mathbf{K}) \subseteq \mathcal{M}(\mathbf{X}'\mathbf{X})$. *Let* $\mathbf{G}$ *be a g-inverse of* $\mathbf{X}'\mathbf{X}$. *Then* $\mathcal{M}(\mathbf{K}') = \mathcal{M}(\mathbf{K}'\,\mathbf{G}\,\mathbf{K})$.

Proof. Clearly $\mathcal{M}(\mathbf{K}'\,\mathbf{G}\,\mathbf{K}) \subseteq \mathcal{M}(\mathbf{K}')$. Let $\mathbf{G}$ be a g-inverse of $\mathbf{X}'\mathbf{X}$. Since $\mathcal{M}(\mathbf{K}) \subseteq \mathcal{M}(\mathbf{X}'\mathbf{X})$ we can write $\mathbf{K} = (\mathbf{X}'\mathbf{X})\,\mathbf{D}$ for some matrix $\mathbf{D}$. Therefore the rank of $\mathbf{K}'\,\mathbf{G}\,\mathbf{K}$ is same as the rank of $\mathbf{D}'(\mathbf{X}'\mathbf{X})\,\mathbf{D}$, which in turn equals the rank of $\mathbf{D}'\,\mathbf{X}'$. Thus we have

$$\text{rank of } (\mathbf{K}'\,\mathbf{G}\,\mathbf{K}) \quad = \quad \text{rank of } (\mathbf{D}'\,\mathbf{X}')$$

$$\geq \text{ rank of } (\mathbf{K}'). \tag{3.17}$$

Since the other inequality always holds, we have rank of $(\mathbf{K}'\,\mathbf{G}\,\mathbf{K})$ equals rank of $\mathbf{K}'$. This completes the proof of the lemma.

ACKNOWLEDGMENT

Research partially supported by the U. S. Army research office grant number DAAH04-96-1-0070. The United States Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon.

REFERENCES

Bellman, R. (1970). *Introduction to Matrix Analysis* (2nd ed.). New York: McGraw-Hill.

Nair, K. R. (1956). "A Note on the Estimation of Mean Change". *Journal of the Indian Society of Agricultural Statistics*, 8, 122-124.

Olkin, I. (1992). "A Matrix Formulation on How Deviant an Observation Can Be". *The American Statistician*, 46, 205-209.

Rao, C. R. (1973). *Linear Statistical Inference and Its Application* (2nd ed.). New York: John Wiley.

Samuelson, P. A. (1968). "How Deviant Can You Be?" *Journal of the American Statistical Association*, 63, 1522-1525.

Searle, S. R. (1982). *Matrix Algebra Useful For Statistics*, New York: John Wiley.

Seber, G. A. F. (1977). *Linear Regression Analysis*, New York: John Wiley.

Thomson, G . W. (1955). "Bounds for the Ratio of Range to Standard Deviation". *Biometrika*, 42, 268-269.

ELSEVIER

Journal of Statistical Planning and Inference 76 (1999) 145–161

journal of statistical planning and inference

On eliminating the asymptotic bias in the quasi-least squares estimate of the correlation parameter

N. Rao Chaganty [a,*], Justine Shults [b]

[a] *Department of Mathematics and Statistics, Old Dominion University, Norfolk, VA 23529-0077, USA*
[b] *Center for Pediatric Research, Eastern Virginia Medical School, Norfolk, VA 23510-1001, USA*

Received 15 August 1997; accepted 7 August 1998

Abstract

In a recent paper, Chaganty (1997, J. Statist. Plann. Inference 63, 39–54) introduced the method of quasi-least squares (QLS) for estimating the regression, correlation and scale parameters in longitudinal data analysis problems. The QLS estimates of the regression and scale parameters are consistent even if the working correlation structure is misspecified. The estimate of the correlation parameter, however, is asymptotically biased. In this paper, we present modified (C-QLS) estimates of the correlation parameter for the following working correlation structures that are appropriate for the analysis of balanced and equally spaced longitudinal data: the unstructured matrix, for which the C-QLS estimate is a positive definite, consistent correlation matrix; and the exchangeable, tridiagonal, and autoregressive structures, for which the C-QLS estimates are feasible, consistent and robust against misspecification. We also present feasible and consistent C-QLS estimates for two structures appropriate for the analysis of unbalanced and unequally spaced longitudinal data: the Markov and generalized Markov working correlation structures that were discussed by Núñez-Anton and Woodworth (1994, Biometrics 50, 445–456) and Shults and Chaganty (1998, Biometrics 54, 1622–1630). We then present an improved consistent estimate of the scale parameter. Finally, examples are given to contrast the C-QLS estimates with estimates obtained using the widely used generalized estimating equation (GEE) approach.

AMS classifications: primary 62J12; 62F10; 62F12; secondary 15A23

Keywords: GEE; Generalized least squares; Longitudinal data; Positive definite matrix; Quasi-least squares; Repeated measures

1. Introduction

We consider longitudinal data that can be described as follows. Let $Y_i = (y_{i1}, \ldots, y_{in})'$ be a vector of repeated measurements taken on the ith subject; associated with each

* Corresponding author. Tel.: +1 757 683 3897; fax: +1 757 683 3885; e-mail: nrc@math.odu.edu.

0378-3758/99/$ – see front matter
PII: S0378-3758(98)00180-3

measurement y_{ij} is a vector of covariates $x_{ij}=(x_{ij1},\ldots,x_{ijp})'$; $1\leqslant j\leqslant n$, $1\leqslant i\leqslant m$. The Y_i's are uncorrelated with an unspecified distribution that satisfies $E(y_{ij})=\mu_{ij}$ and $\text{var}(y_{ij})=\phi v(\mu_{ij})$. The variance function v is assumed to be known but $\phi>0$ may be a known constant or an unknown scale parameter. We also assume that there is an invertible function g, known as the link function, such that $\mu_{ij}=\mu_{ij}(\beta)=g^{-1}(x_{ij}'\beta)$, where $\beta\in\mathcal{R}^p$ is a vector of unknown regression coefficients. The correlation between the repeated measurements on each subject is modelled by a working correlation matrix $R(\alpha)$, which is a function of the vector $\alpha=(\alpha_1,\ldots,\alpha_q)'$. The set of feasible values of α is a subset $\mathcal{S}$ of $\mathcal{R}^q$ such that $R(\alpha)$ is a positive-definite matrix for $\alpha\in\mathcal{S}$. The vector α is considered to be an unknown nuisance parameter that must be eliminated to estimate β, the main parameter of interest.

There are numerous papers in the literature concerning estimation of β and α that use the method of generalized estimating equations (GEE), introduced by Liang and Zeger (1986). In this paper, we primarily focus on an alternative method of estimation known as quasi-least squares (QLS) that was described in Chaganty (1997) and Shults and Chaganty (1998). Both the GEE and the QLS methods use the same estimating equation for β and both methods yield a consistent estimate for β. Furthermore, as $m\to\infty$, the asymptotic relative efficiency of the QLS estimate of β with respect to the corresponding GEE estimate is 1. The two methods differ however, in estimation of the working correlation parameter α. The QLS method estimates α by minimizing the generalized error sum of squares (see Eq. (2.1)), whereas the GEE method uses a moment estimate for the correlation parameter α.

The QLS estimate of the correlation parameter α is asymptotically biased. In this paper, we eliminate this asymptotic bias by modifying the QLS estimate of the correlation parameter using continuous and one-to-one transformations that depend on the working correlation matrix. Our goal in eliminating the bias of the QLS correlation parameter estimate is to allow for consistent estimation of the standard errors of the regression parameter estimate. Consider first the situation where we have an equal number of measurements on each subject that are observed at equally spaced time points. Suppose that the working correlation is totally unspecified. In this case the modified QLS (C-QLS) estimate of the correlation matrix is positive definite and consistent. Next, suppose that a structured correlation matrix can be used to describe the pattern of correlation among observations collected on each subject and that reasonable candidates for the correlation structure include the AR(1), equicorrelated, and the tridiagonal. Each of these structures depends on a single parameter ρ, which represents the correlation between two adjacent observations collected on a subject. (We shall use α to denote the working correlation parameter and ρ to denote the true correlation parameter.) In this situation, we recommend using a two-stage procedure to estimate the regression and correlation parameters. The first stage uses the AR(1) structure as the working structure; it yields a C-QLS estimate $\hat{\rho}_m$ of the correlation parameter ρ that is not only feasible and consistent, but is also robust against misspecification of the correlation structure among the AR(1), equicorrelated, and tridiagonal working correlation structures. In the second stage we obtain the final C-QLS estimate of the regression

parameter by solving the estimating equation for β using $\hat{\rho}_m$ and the most appropriate working structure for the data being analyzed. This will ensure that we do not lose any efficiency in estimating the regression parameter.

Consider now the situation in which the observations on each subject are unbalanced and unequally spaced and the correlation between measurements on each subject depends on their separation in time. The intra-subject correlation may also stabilize over time, in the sense that two successive measurements taken on a subject later during a study will be more highly correlated than if they are collected earlier. Two correlation models that are appropriate in these situations are the Markov and the generalized Markov (see Núñez-Anton and Woodworth, 1994). As was discussed in Shults and Chaganty (1998), the method of GEE may yield infeasible estimates of the correlation parameter for the Markov structure and cannot easily be applied for the generalized Markov structure. In this paper, we derive a consistent C-QLS estimate of the correlation parameter for the Markov and generalized Markov structures under an assumption that the correlation model has been correctly specified. The C-QLS estimate of the regression parameter β is obtained by solving the estimating equation for β using the C-QLS estimate of the correlation parameter ρ.

The organization of this paper is as follows: In Section 2 we briefly describe the method of QLS and give an expression for the asymptotic bias of the correlation parameter estimate. In Section 3 we obtain consistent estimates of the true correlation parameter using continuous and one-to-one transformations on the QLS estimate of the working correlation parameter for several correlation models: the unstructured (Section 3.1); the AR(1), tridiagonal, and equicorrelated (Section 3.2); and the Markov (Section 3.3) and generalized Markov (Section 3.4). In Section 4 we propose an improved consistent estimate of the scale parameter ϕ. We then apply GEE and the C-QLS approach with the modified regression, correlation and scale parameter estimates in analyses of equally spaced and balanced dental data (Section 5.1) and unequally spaced and unbalanced audiology data (Section 5.2). Finally, the appendix contains the proof of consistency of the scale parameter estimate.

2. Quasi-least squares

Here, we briefly describe the method of quasi-least squares when the data comprise equal numbers of measurements (balanced observations) that are collected at equally spaced time points on each of a group of independent subjects. For a more detailed description see Chaganty (1997) and Shults and Chaganty (1998).

Let $Z_i(\beta) = A_i^{-1/2}(\beta)(Y_i - \mu_i(\beta))$, where $\mu_i(\beta) = (\mu_{i1}(\beta), \ldots, \mu_{in}(\beta))'$ and $A_i(\beta) = \operatorname{diag}(v(\mu_{i1}(\beta)), \ldots, v(\mu_{in}(\beta)))$ be the vector of means and diagonal matrix of variances, respectively; $1 \leqslant i \leqslant m$. The method of QLS obtains estimates by partially minimizing the generalized error sum of squares

$$Q(\beta, R(\alpha)) = \sum_{i=1}^{m} Z_i'(\beta)\, R^{-1}(\alpha)\, Z_i(\beta) \tag{2.1}$$

with respect to $\beta \in \mathscr{R}^p$ and $\alpha \in \mathscr{S}$. Note that the quadratic form Eq. (2.1) not only depends on β and α, but also on the structure of the correlation matrix $R(\alpha)$. The estimating equations obtained by taking partial derivatives of Eq. (2.1) with respect to β and α, are

$$\sum_{i=1}^{m} D_i'(\beta) A_i^{-1/2}(\beta)\, R^{-1}(\alpha)\, Z_i(\beta) = 0 \tag{2.2}$$

and

$$\sum_{i=1}^{m} Z_i'(\beta) \frac{\partial R^{-1}(\alpha)}{\partial \alpha_j} Z_i(\beta) = 0, \quad 1 \leqslant j \leqslant q, \tag{2.3}$$

where $D_i(\beta) = \partial \mu_i / \partial \beta'$. The QLS estimates $\hat{\beta}_m$ and $\hat{\alpha}_m$ of β and α are the solutions of the two Eqs. (2.2) and (2.3). Let $\bar{R}$ be the true unknown correlation structure between the repeated measurements on each subject. Under some conditions, appealing to the weak law of large numbers, Chaganty (1997) has shown that

$$\hat{\beta}_m \to \beta \quad \text{and} \quad \hat{\alpha}_m \to \alpha - \mathscr{I}_{22}^{-1}(\alpha) a(\alpha) \tag{2.4}$$

in probability as $m \to \infty$, where

$$a(\alpha) = \left[\operatorname{tr}\left(\frac{\partial R^{-1}(\alpha)}{\partial \alpha_j} \bar{R} \right) \right]_{q \times 1} \quad \text{and} \quad \mathscr{I}_{22}(\alpha) = \left[\operatorname{tr}\left(\frac{\partial^2 R^{-1}(\alpha)}{\partial \alpha_j \, \partial \alpha_k} \bar{R} \right) \right]_{q \times q}. \tag{2.5}$$

(The $-$ sign in Eq. (2.4) was incorrectly written as $+$ in Chaganty (1997)). It follows from Eq. (2.4) that $\hat{\beta}_m$ is a consistent estimate of β, even if the working correlation is misspecified. But $\hat{\alpha}_m$ has an asymptotic bias given by $\mathscr{I}_{22}^{-1}(\alpha) a(\alpha)$, which depends both on the working and the true correlation matrices.

3. Consistent estimate of the true correlation parameter

Here, we obtain continuous and one-to-one transformations of the QLS estimate of the working correlation parameter to obtain a consistent estimate of the true correlation parameter. These transformations depend on the working correlation structure that is most appropriate for our data. We first consider longitudinal data that are balanced and equally spaced in time. We derive transformations in a closed form for the unstructured correlation matrix (Section 3.1) and for the AR(1), equicorrelated, and tridiagonal structures (Section 3.2). We next derive bias-eliminating transformations for two structures appropriate for the analysis of unbalanced and unequally spaced data – the Markov (Section 3.3) and the generalized Markov (Section 3.4). These transformations are not in a closed form, but the C-QLS estimates of the correlation parameters can be obtained numerically.

3.1. Unstructured correlation matrix

Suppose that the working correlation matrix, $R(\alpha)=\bar{R}$ is unstructured. Here the feasible set $\mathscr{S}_u$ is the class of all positive-definite correlation matrices. Let $\hat{\beta}_{um}$ and $\hat{R}_m$ be the QLS estimates of β and $\bar{R}$, respectively. Since the QLS estimate $\hat{R}_m$ is asymptotically biased, we will use a transformation on $\hat{R}_m$ to obtain a consistent estimate. From the results of Whittle (1958) and Olkin and Pratt (1958) we know that every positive definite matrix Σ admits a unique decomposition

$$\Sigma=\tilde{R}\Lambda\tilde{R}, \tag{3.1}$$

where Λ is a diagonal matrix of positive elements and $\tilde{R}$ is a correlation matrix. Clearly, decomposition (3.1) also holds for the subclass of positive definite correlation matrices. The function $f_u: R\ (=\tilde{R}\Lambda\tilde{R})\to\tilde{R}$ is a continuous, one-to-one and onto mapping from $\mathscr{S}_u$ to $\tilde{\mathscr{S}}_u$, where $\tilde{\mathscr{S}}_u=\{\tilde{R}\in\mathscr{S}_u: v=(\tilde{R}\circ\tilde{R})^{-1}e>0\}$. Here, e is a vector of ones and $\circ$ denotes the Hadamard product. Furthermore $f_u(\phi R)=f(R)$ for all $\phi>0$. It is easy to verify that for $\tilde{R}\in\tilde{\mathscr{S}}_u$, $f_u^{-1}(\tilde{R})=\tilde{R}\Lambda\tilde{R}=R$, where $\Lambda=\mathrm{diag}(v)$. For $n=2$ we have $\tilde{\mathscr{S}}_u=\mathscr{S}_u$ and for $n>2$, the set $\tilde{\mathscr{S}}_u$ is a proper open subset of $\mathscr{S}_u$. See Olkin and Pratt (1958) (p. 233) for an example of a correlation matrix that is in $\mathscr{S}_u$ but not in $\tilde{\mathscr{S}}_u$.

If the working correlation is unstructured, we can obtain a bias corrected estimate $\hat{R}_{cm}$, of the true correlation matrix using the following three steps:

Step 1: Assume that the working correlation matrix is unstructured and compute the QLS estimates $\hat{\beta}_{um}$ and $\hat{R}_m$. See Chaganty (1997) (Example 4.4) for computational details.

Step 2: Compute $\hat{\boldsymbol{Z}}_{um}=(1/m)\sum_{i=1}^{m}Z_i(\hat{\beta}_{um})Z_i(\hat{\beta}_{um})'$ and $\hat{v}_m=(\hat{R}_m\circ\hat{R}_m)^{-1}e$, where $\circ$ denotes the Hadamard product.

Step 3: Obtain the modified estimate of the correlation matrix,

$$\hat{R}_{cm}=\begin{cases}\hat{R}_{um}=f_u^{-1}(\hat{R}_m)=\hat{R}_m\,\mathrm{diag}(\hat{v}_m)\,\hat{R}_m & \text{if } \hat{v}_m>0 \text{ (i.e. } \hat{R}_m\in\tilde{\mathscr{S}}_u),\\ \hat{R}_{sm}=(\mathrm{diag}(\hat{\boldsymbol{Z}}_{um}))^{-1/2}\hat{\boldsymbol{Z}}_{um}(\mathrm{diag}(\hat{\boldsymbol{Z}}_{um}))^{-1/2} & \text{otherwise.}\end{cases} \tag{3.2}$$

We will now establish that $\hat{R}_{cm}$ is a consistent estimate of $\bar{R}$. From Example 4.4 in Chaganty (1997) we know that $\hat{\boldsymbol{Z}}_{um}$ can be written as

$$\hat{\boldsymbol{Z}}_{um}=\hat{R}_m\Lambda_m\hat{R}_m, \tag{3.3}$$

where Λ_m is a diagonal matrix of positive elements. Since $\hat{\boldsymbol{Z}}_{um}\to\phi\bar{R}$ almost surely, as $m\to\infty$, and decompositions (3.1) and (3.3) are unique, it is easy to see that $f_u(\phi\bar{R})=f_u(\bar{R})=\bar{\bar{R}}$, where $\bar{\bar{R}}=\lim_{m\to\infty}\hat{R}_m$. Also for sufficiently large m, $\hat{R}_m\in\tilde{\mathscr{S}}_u$, since $\tilde{\mathscr{S}}_u$ is an open set and $\bar{\bar{R}}\in\tilde{\mathscr{S}}_u$. Therefore,

$$\hat{R}_{um}=f_u^{-1}(\hat{R}_m)\to f_u^{-1}(\bar{\bar{R}})=\bar{R} \tag{3.4}$$

in probability as $m \to \infty$. Clearly $\hat{R}_{sm}$ is a consistent estimate of $\bar{R}$. Therefore, the modified estimate $\hat{R}_{cm}$ is also a consistent estimate of the true correlation matrix $\bar{R}$.

Remark 3.1. From the above discussion it is clear that $\hat{R}_{cm} = f_u^{-1}(\hat{R}_m)$ (i.e. $\hat{R}_m \in \tilde{\mathscr{S}}_u$) almost surely for sufficiently large values of m. In some longitudinal data analysis problems it is possible that $\hat{R}_m \notin \tilde{\mathscr{S}}_u$. We should view this outcome as an indication that our assumption, $\text{var}(y_{ij}) = \phi v(\mu_{ij})$ may be incorrect and that the correct specification might instead be $\text{var}(y_{ij}) = \phi_j v(\mu_{ij})$. In either case, $\hat{R}_{cm} = \hat{R}_{sm}$ is a consistent estimate of $\bar{R}$.

Remark 3.2. Note that the unstructured working correlation can also be used to analyze longitudinal outcomes that are not equally spaced; however, the timings of the measurements should be the same for all the subjects.

3.2. Structured correlation matrix: Balanced and equally spaced data

When analyzing balanced and equally spaced data, use of a structured correlation matrix is often preferable to the use of an unstructured matrix. One important advantage afforded by fitting a structured matrix is that it will allow for parsimonious modelling of the regression and correlation parameters.

The bias correcting technique that was used in Section 3.1 for unstructured correlation can be stated more formally in the following theorem for structured correlation matrices.

Theorem 3.2. *Let β, α, ρ and ϕ be fixed. Let $R(\alpha)$ be the working correlation structure, and assume that the true correlation $\bar{R}(\rho)$ is also structured, where α and ρ are vectors in $\mathscr{S}$ which is a subset of $\mathscr{R}^q$. Suppose that the solution of the equation*

$$b(\alpha, \rho) = \left[\text{tr}\left(\frac{\partial R^{-1}(\alpha)}{\partial \alpha_j} \bar{R}(\rho) \right) \right]_{q \times 1} = 0 \tag{3.5}$$

is given by $\alpha = f(\rho)$, that is, $b(f(\rho), \rho) = 0$ or equivalently $b(\alpha, f^{-1}(\alpha)) = 0$, where f is a continuous and one-to-one function. If $\hat{\alpha}_m$ is the QLS estimate of α then the C-QLS estimate $\hat{\rho}_m = f^{-1}(\hat{\alpha}_m)$ is a consistent estimate of ρ.

Proof. Let $\hat{\beta}_m$ and $\hat{\alpha}_m$ be the QLS estimates of β and α, respectively, when the working correlation structure is $R(\alpha)$. Note that $\hat{\alpha}_m$ is the solution of the equation

$$\left[\text{tr}\left(\frac{\partial R^{-1}(\alpha)}{\partial \alpha_j} \hat{\boldsymbol{Z}}_m \right) \right]_{q \times 1} = 0, \tag{3.6}$$

where $\hat{\boldsymbol{Z}}_m = (1/m) \sum_{i=1}^m Z_i(\hat{\beta}_m) Z_i(\hat{\beta}_m)'$. Since $\hat{\beta}_m \to \beta$, we have $\hat{\boldsymbol{Z}}_m \to \phi \bar{R}(\rho)$ and therefore, $\hat{\alpha}_m \to f(\rho)$ in probability, as $m \to \infty$. Hence, $\hat{\rho}_m = f^{-1}(\hat{\alpha}_m) \to \rho$ in probability as $m \to \infty$. This completes the proof of the theorem. □

Remark 3.3. An alternative view of our bias correcting technique can be summarized as follows. Note that the expected value of $\boldsymbol{Z}_m = (1/m)\sum_{i=1}^{m} Z_i(\beta)\, Z_i(\beta)'$ equals $\phi\bar{R}(\rho)$. If we modify the estimating Eq. (2.3) so as to make it unbiased we obtain the equation

$$\left[\operatorname{tr}\left(\frac{\partial R^{-1}(\alpha)}{\partial \alpha_j}(\boldsymbol{Z}_m - \phi\bar{R}(\rho))\right)\right]_{q\times 1} = 0. \tag{3.7}$$

To get a consistent estimate of the true correlation parameter ρ, we could directly solve the unbiased estimating Eq. (3.7) replacing β with $\hat{\beta}_m$, that is, replacing $\boldsymbol{Z}_m$ with $\hat{\boldsymbol{Z}}_m$. But there are two drawbacks to this approach: direct solution of Eq. (3.7) requires estimation of ϕ; and, for some common working correlation structures, even if the working correlation is correctly specified, a feasible solution may not exist. Our method of estimation overcomes these drawbacks. Note that the estimates $\hat{\beta}_m$, $\hat{\alpha}_m$ and $\hat{\rho}_m$ satisfy

$$\left[\operatorname{tr}\left(\left.\frac{\partial R^{-1}(\alpha)}{\partial \alpha_j}\right|_{\alpha=\hat{\alpha}_m} \hat{\boldsymbol{Z}}_m\right)\right]_{q\times 1} = 0 \tag{3.8}$$

and

$$\left[\operatorname{tr}\left(\left.\frac{\partial R^{-1}(\alpha)}{\partial \alpha_j}\right|_{\alpha=\hat{\alpha}_m} \bar{R}(\hat{\rho}_m)\right)\right]_{q\times 1} = 0. \tag{3.9}$$

Therefore, we have

$$\left[\operatorname{tr}\left(\left.\frac{\partial R^{-1}(\alpha)}{\partial \alpha_j}\right|_{\alpha=\hat{\alpha}_m} (\hat{\boldsymbol{Z}}_m - \phi\bar{R}(\hat{\rho}_m))\right)\right]_{q\times 1} = 0 \quad \forall \phi > 0. \tag{3.10}$$

Thus, the C-QLS method gives a solution to the unbiased estimating Eq. (3.7) that does not depend on ϕ. Moreover, the estimate $\hat{\rho}_m$ of the true correlation parameter exists, is feasible, unique, and easy to compute. The standard conditions required to establish consistency are satisfied so that $\hat{\rho}_m$ is indeed consistent.

Remark 3.4. Theorem 3.2 requires the specification of the working and the true underlying correlation structure of our data. It will therefore be useful when the working correlation is correctly specified, or when the function $f(\rho)$ in Theorem 3.2 does not depend on the structure of the true correlation matrix $\bar{R}$. The latter situation occurs in the analysis of balanced and equally spaced data if we choose an appropriate working correlation structure. For unbalanced and unequally spaced correlated data we will assume that the former is true, that is, that we have correctly specified the working correlation matrix. However, the correlation model that we consider in this paper is extremely flexible and thus is appropriate for data that have a wide range of characteristics that are typical for unbalanced and unequally spaced longitudinal data. Because of this, and in the absence of appropriate alternative correlation structures, the

assumption that we have correctly specified the working correlation model will be reasonable in most analyses of unbalanced and unequally spaced longitudinal outcomes.

In many practical situations when the longitudinal data are balanced and are observed at equally spaced time intervals, the correlation parameter α is a real variable and the useful and popular working correlation structures are: (i) identity ($R^{(1)}$); (ii) equicorrelated ($R^{(2)}(\alpha)$); here all the off-diagonal elements equal α, (iii) AR(1) ($R^{(3)}(\alpha)$); the (i,j) element for this structure is given by $\alpha^{|i-j|}$ and (iv) tridiagonal ($R^{(4)}(\alpha)$); the elements just above and below the diagonal equal α and the others are zero. We observe that if the working correlation matrix $R(\alpha)$ is AR(1) then the solution of Eq. (3.5) is

$$\alpha = f_a(\rho) = \begin{cases} \dfrac{1-\sqrt{1-\rho^2}}{\rho} & \text{if } \rho \neq 0, \\ 0 & \text{if } \rho = 0, \end{cases} \tag{3.11}$$

when the true correlation structure $\bar{R}(\rho) = R^{(j)}$, $j = 1,2,3,4$. We exploit this important observation to obtain a consistent and robust estimate of the true correlation parameter ρ. Note that for AR(1) working correlation structure, the set $\mathcal{S}_a$ of feasible values of α is the open interval $(-1,1)$ and the function $f_a(\rho)$ is a continuous, one-to-one and onto mapping from $\mathcal{S}_a$ to $\mathcal{S}_a$. The inverse mapping is

$$f_a^{-1}(\alpha) = \rho = \frac{2\alpha}{1+\alpha^2}. \tag{3.12}$$

Let $\hat{\beta}_{am}$ and $\hat{\alpha}_{am}$ be the QLS estimates of β and α, respectively, when the working correlation has AR(1) structure. Then it follows from Theorem 3.2 that

$$\hat{\rho}_{am} = f_a^{-1}(\hat{\alpha}_{am}) \to \rho \tag{3.13}$$

in probability as $m \to \infty$, when $\bar{R}(\rho) = R^{(j)}$ for $j = 2,3,4$. We can easily check that if $\alpha = 0$ then $\hat{\alpha}_{am}$, as well as $\hat{\rho}_{am}$, both converge to 0 in probability as $m \to \infty$, when the true correlation matrix is $R^{(1)}$, the identity matrix. Therefore, the estimate $\hat{\rho}_{am}$ of the true correlation parameter ρ is consistent, feasible, and robust against misspecification among the four most widely used correlation models for analyzing balanced and equally spaced longitudinal data.

3.3. *Unbalanced and unequally spaced data: Markov structure*

Suppose that the longitudinal data are unbalanced and that n_i measurements are made on subject i at times $0 < t_{i1} < t_{i2} < \cdots < t_{in_i}$; $1 \leqslant i \leqslant m$. A suitable working correlation for these unequally spaced repeated measurements is the generalized Markov structure

given by

$$R_i(\alpha)=\begin{pmatrix} 1 & \eta^{e_{i2}} & \eta^{e_{i2}+e_{i3}} & \cdots & \eta^{e_{i2}+e_{i3}+\cdots+e_{in_i}} \\ \eta^{e_{i2}} & 1 & \eta^{e_{i3}} & \cdots & \eta^{e_{i3}+e_{i4}+\cdots+e_{in_i}} \\ \eta^{e_{i2}+e_{i3}} & \eta^{e_{i3}} & 1 & \cdots & \eta^{e_{i4}+e_{i5}+\cdots+e_{in_i}} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \eta^{e_{i2}+e_{i3}+\cdots+e_{i(n_i-1)}} & \cdot & \cdot & \cdots & \eta^{e_{in_i}} \\ \eta^{e_{i2}+e_{i3}+\cdots+e_{in_i}} & \eta^{e_{i3}+e_{i4}+\cdots+e_{in_i}} & \cdot & \cdot\cdot\eta^{e_{in_i}} & 1 \end{pmatrix}, \tag{3.14}$$

where $\alpha=(\alpha_1=\eta,\ \alpha_2=\lambda)'$ and the e_{ik}'s are functions of the parameter λ defined as follows:

$$e_{ik}(\lambda)=\begin{cases} \dfrac{[t_{ik}^{\lambda}-t_{i(k-1)}^{\lambda}]}{\lambda} & \text{if } \lambda\neq 0, \\ \log(t_{ik})-\log(t_{i(k-1)}) & \text{if } \lambda=0, \end{cases} \tag{3.15}$$

for $2\leqslant k\leqslant n_i;\ 1\leqslant i\leqslant m$. The feasible range for the parameter λ is $(-\infty,\infty)$ and η is restricted to $(0,1)$. We will discuss the bias correction for this structure in Section 3.4, but first we consider the Markov correlation model, the important special case of the generalized Markov structure when $\lambda=1$. The Markov structure is appropriate when the correlation between unequally spaced measurements collected on a subject decreases with increasing separation in time. Here, $e_{ik}=[t_{ik}-t_{i(k-1)}];\ 2\leqslant k\leqslant n_i;\ 1\leqslant i\leqslant m$. The correlation parameter $\alpha=\eta$ is a real variable and is restricted to the interval $(0,1)$. Let us assume that the working correlation is the Markov structure and it is correctly specified, that is, the true correlation structure $\bar{R}_i$ is also given by Eq. (3.14) and $\lambda=1$. The bias correcting equation in this case is

$$\begin{aligned} b(\alpha,\rho)&=\sum_{i=1}^{m}\operatorname{tr}\left(\frac{\partial R_i^{-1}(\alpha)}{\partial\alpha}\bar{R}_i(\rho)\right) \\ &=2\sum_{i=1}^{m}\sum_{k=2}^{n_i}\frac{2e_{ik}\alpha^{2e_{ik}-1}-\rho^{e_{ik}}e_{ik}[\alpha^{e_{ik}-1}+\alpha^{3e_{ik}-1}]}{[1-\alpha^{2e_{ik}}]^2} \\ &=0. \end{aligned} \tag{3.16}$$

Note that Eq. (3.16) reduces to Eq. (3.12) when the data are balanced and equally spaced, that is, $n_i=n$ and $e_{ik}=1$ for all i and k. Let $\hat{\alpha}_m$ be the QLS estimate. The bias corrected estimate is then obtained by solving the equation

$$b(\hat{\alpha}_m,\rho)=0 \tag{3.17}$$

for ρ. That is, given $\hat{\alpha}_m$ the bias corrected estimate $\hat{\rho}_m$ satisfies the equation

$$\sum_{i=1}^{m}\sum_{k=2}^{n_i}\frac{2e_{ik}\hat{\alpha}_m^{2e_{ik}-1}}{[1-\hat{\alpha}_m^{2e_{ik}}]^2}=\sum_{i=1}^{m}\sum_{k=2}^{n_i}\frac{\hat{\rho}_m^{e_{ik}}\ e_{ik}[\hat{\alpha}_m^{e_{ik}-1}+\hat{\alpha}_m^{3e_{ik}-1}]}{[1-\hat{\alpha}_m^{2e_{ik}}]^2}, \tag{3.18}$$

where $e_{ik}=[t_{ik}-t_{i(k-1)}]$; $2\leqslant k\leqslant n_i$; $1\leqslant i\leqslant m$. That there is a unique solution $\hat{\rho}_m$ for Eq. (3.18) that lies in the interval (0, 1) can be shown as follows. Let us denote the l.h.s. of Eq. (3.18) by c. Let

$$h(\rho)=\sum_{i=1}^{m}\sum_{k=2}^{n_i}\frac{\rho^{e_{ik}}\ e_{ik}[\hat{\alpha}_m^{e_{ik}-1}+\hat{\alpha}_m^{3e_{ik}-1}]}{[1-\hat{\alpha}_m^{2e_{ik}}]^2}. \tag{3.19}$$

Clearly, the function $h(\rho)$ is a continuous and increasing function of ρ, since $\hat{\alpha}_m\in(0,1)$ and the e_{ik}'s are all positive. Also $h(0)=0$ and we can easily verify that $h(1)>c$. By the mean value theorem we conclude that there exists a unique $\hat{\rho}_m\in(0,1)$ such that $h(\hat{\rho}_m)=c$. The estimate $\hat{\rho}_m$ can be computed numerically using the bisection method.

3.4. *Unbalanced and unequally spaced data: Generalized Markov structure*

The Markov structure is widely used when the intra-subject correlation of measurements decreases with increasing separation in time. However, this structure may force the intra-subject correlations to decrease too rapidly with increasing separation in time and it does not take into account the actual timings of measurements in the study. When the primary outcome of interest stabilizes over time within subjects, two successive measurements collected on a subject later during the study will be more highly correlated than if they were collected earlier, so that the correlation between these measurements will depend on their time of occurence in the study. The generalized Markov structure generalizes the Markov model so that the decrease in intra-subject correlations may be dampened (or accelerated) with increasing separation in time. It also allows for stabilization in the outcome variable over time; see Núñez-Anton and Woodworth (1994) and Shults and Chaganty (1998) for a detailed discussion of these structures.

Suppose that the true and working correlation structures both are generalized Markov. Let $\alpha=(\eta,\lambda)$ and $\rho=(\bar{\eta},\bar{\lambda})$. For convenience of notation we will supress the argument λ and write e_{ik} for $e_{ik}(\lambda)$ and $\bar{e}_{ik}=e_{ik}(\bar{\lambda})$. Since we have unequal number of observations on each subject the appropriate bias correcting equations analogous to Eq. (3.5) are the following two equations:

$$b_1(\alpha,\rho)=\sum_{i=1}^{m}\operatorname{tr}\left(\frac{\partial R_i^{-1}(\alpha)}{\partial\eta}\bar{R}_i(\rho)\right)=0 \tag{3.20}$$

and

$$b_2(\alpha,\rho)=\sum_{i=1}^{m}\operatorname{tr}\left(\frac{\partial R_i^{-1}(\alpha)}{\partial\lambda}\bar{R}_i(\rho)\right)=0. \tag{3.21}$$

Let $\hat{\alpha}_m = (\eta_m, \lambda_m)$ be the QLS estimate of $\alpha = (\eta, \lambda)$. The C-QLS estimate $\hat{\rho}_{gm} = (\bar{\eta}_m, \bar{\lambda}_m)$ is obtained by solving for ρ simultaneously the equations $b_1(\hat{\alpha}_m, \rho) = 0$ and $b_2(\hat{\alpha}_m, \rho) = 0$, which after simplification reduce to the following two equations:

$$\sum_{i=1}^{m} \sum_{k=2}^{n_i} \frac{\hat{e}_{ik} \eta_m^{\hat{e}_{ik}-1} [2\eta_m^{\hat{e}_{ik}} - \bar{\eta}^{\hat{e}_{ik}}(1 + \eta_m^{2\hat{e}_{ik}})]}{(1 - \eta_m^{2\hat{e}_{ik}})^2} = 0 \tag{3.22}$$

and

$$\sum_{i=1}^{m} \sum_{k=2}^{n_i} \frac{\eta_m^{\hat{e}_{ik}} \frac{\partial \hat{e}_{ik}}{\partial \lambda_m} [2\eta_m^{\hat{e}_{ik}} - \bar{\eta}^{\hat{e}_{ik}}(1 + \eta_m^{2\hat{e}_{ik}})]}{(1 - \eta_m^{2\hat{e}_{ik}})^2} = 0, \tag{3.23}$$

where $\hat{e}_{ik} = e_{ik}(\lambda_m)$ and $\dfrac{\partial e_{ik}}{\partial \lambda_m} = \dfrac{t_{ik}^{\lambda_m}(\log(t_{ik})) - t_{i(k-1)}^{\lambda_m}(\log(t_{i(k-1)}))}{\lambda_m} - \dfrac{t_{ik}^{\lambda_m} - t_{i(k-1)}^{\lambda_m}}{\lambda_m^2}$.

We used the MATLAB Optimization Toolbox routine 'CONSTR' to obtain the QLS estimate $\hat{\alpha}_m = (\eta_m, \lambda_m)$ and the MATLAB routine 'FSOLVE' to solve Eqs. (3.22) and (3.23) to obtain C-QLS estimate $\hat{\rho}_{gm} = (\bar{\eta}_m, \bar{\lambda}_m)$ in the example discussed in Section 5.2.

4. Consistent estimate of the scale parameter

Now, suppose that the scale parameter ϕ is unknown. In this section we will obtain a consistent estimate of ϕ for working correlation structures that are appropriate for balanced and equally spaced observations and also for unbalanced and unequally spaced longitudinal data.

4.1. Balanced and equally spaced data

Suppose that the longitudinal data are balanced and are observed at equally spaced time points. In Theorems 4.1 and 4.2, we obtain a consistent estimate of ϕ when the working correlation is unstructured and structured, respectively.

Theorem 4.1. *Let β and ϕ be fixed. Let $Q(\beta, R(\alpha))$ be as defined in* Eq. (2.1). *Let $\bar{R}$ be the true unknown correlation matrix. Assume that the working correlation structure is totally unspecified, that is, $R(\alpha) = \bar{R}$. Let $(\hat{\beta}_{um}, \hat{R}_m)$ be the solution of the* Eqs. (2.2) *and* (2.3). *Let $\hat{R}_{cm}$ be as defined in* Eq. (3.2). *Assume that the conditions of Theorem* 5.1 *in Chaganty* (1997) *hold. Then*

$$\frac{Q(\hat{\beta}_{um}, \hat{R}_{cm})}{mn} \to \phi \tag{4.1}$$

in probability as $m \to \infty$.

Theorem 4.2. *Let β, α and ϕ be fixed. Let $Q(\beta, R(\alpha))$ be as defined in* Eq. (2.1). *Assume that the working correlation structure is $AR(1)$, that is, $R(\alpha) = R^{(3)}(\alpha)$. Let $(\hat{\beta}_{am}, \hat{\alpha}_{am})$ be the solution of the* Eqs. (2.2) *and* (2.3). *Let $\hat{\rho}_{am}$ be as defined in*

Eq. (3.13). *Assume that the conditions of Theorem* 5.1 *in Chaganty* (1997) *hold. Then*

$$\frac{Q(\hat{\beta}_{am}, R^{(3)}(\hat{\rho}_{am}))}{mn} \to \phi \tag{4.2}$$

in probability as $m \to \infty$, *when the true correlation structure* $\bar{R}$, *is any one of the following*: (i) *equicorrelated*, (ii) $AR(1)$ *and* (iii) *tridiagonal. Further, if* $\rho = 0$ *than* Eq. (4.2) *also holds, when* $\bar{R}$ *equals the identity matrix.*

The proof of Theorem 4.2 is given in the appendix, Theorem 4.1 is proved similarly. From the above theorems we can see that a consistent estimate of ϕ is

$$\hat{\phi}_c = \begin{cases} Q(\hat{\beta}_{um}, \hat{R}_{cm})/mn & \text{if } R(\alpha) = \bar{R}, \\ Q(\hat{\beta}_{am}, R^{(3)}(\hat{\rho}_{am}))/mn & \text{if } R(\alpha) = R^{(3)}(\alpha). \end{cases} \tag{4.3}$$

Since the QLS estimates partially minimize the quadratic form (2.1), in practice for small samples, the estimate $\hat{\phi}_c$ will be smaller than the popular consistent estimate of ϕ given by

$$\hat{\phi}_p = \begin{cases} Q(\hat{\beta}_{um}, I)/mn & \text{if } R(\alpha) = \bar{R}, \\ Q(\hat{\beta}_{am}, I)/mn & \text{if } R(\alpha) = R^{(3)}(\alpha), \end{cases} \tag{4.4}$$

where I is the identity matrix. Therefore, a good consistent estimate of ϕ is $\hat{\phi}_g = \min(\hat{\phi}_c, \hat{\phi}_p)$, since it yields shorter width confidence intervals for linear functions of β than $\hat{\phi}_p$ or $\hat{\phi}_c$.

4.2. Unbalanced and unequally spaced data

Suppose that we have n_i observations measured on subject i and that these observations may be unequally spaced in time. Assume that the working correlation structure is correctly specified and that it is Markov (generalized Markov). Let $\hat{\beta}_{gm}$ and $\hat{\rho}_{gm}$ be the C-QLS estimates of the regression parameter β and the true correlation parameter ρ, when the working correlation is Markov (generalized Markov). Let

$$\hat{\phi}_p = \frac{1}{m} \sum_{i=1}^{m} \frac{Z_i'(\hat{\beta}_{gm}) Z_i(\hat{\beta}_{gm})}{n_i} \tag{4.5}$$

and

$$\hat{\phi}_c = \frac{1}{m} \sum_{i=1}^{m} \frac{Z_i'(\hat{\beta}_m) R_i^{-1}(\hat{\rho}_{gm}) Z_i(\hat{\beta}_m)}{n_i}, \tag{4.6}$$

where the correlation matrix R_i is given in Eq. (3.14). A good consistent estimate of ϕ is given by $\hat{\phi}_g = \min(\hat{\phi}_c, \hat{\phi}_p)$.

5. Examples

To contrast the QLS modified regression, correlation and scale parameter estimates described in Sections 3 and 4 with the corresponding GEE estimates, in this section we present the results of two analyses. The first is of balanced, equally spaced data, for which the AR(1) and the unstructured are reasonable candidates for a working correlation model. The second is of unbalanced, unequally spaced data, for which the Markov and generalized Markov structures are appropriate.

5.1. Analysis of balanced and equally spaced data

Here we analyze the longitudinal data displayed in Table 1 of Potthoff and Roy (1964). The data were collected in a dental study of 27 subjects (11 girls and 16 boys). They comprise measurements (y_{ij}'s), in millimeters, from the center of each subjects pituitary to pteryomaxillary fissure recorded at 8, 10, 12 and 14 yr of age. Jennrich and Schluchter (1986) analyzed these data using maximum likelihood procedures to illustrate the use of different covariance structures to model repeated measurements. We fit the following regression model (Model 2 in Jennrich and Schluchter (1986)):

$$\mu_{ij} = \beta_g x_{i1} + \beta_b x_{i2} + \gamma_g x_{i1} * x_{j3} + \gamma_b x_{i2} * x_{j3}, \quad 1 \leqslant j \leqslant 4, \ 1 \leqslant i \leqslant 27, \tag{5.1}$$

where x_{i1}, x_{i2} are indicator variables for the two sexes, girl and boy, respectively. The covariate x_{j3} is the subject's age at the jth measurement time. It takes the values 8, 10, 12 and 14.

Table 1 contains estimates and standard errors for the regression parameters and the estimate of the scale parameter, computed using the GEE and C-QLS methods. The estimates were computed using both the AR(1) and the unstructured (UNSTR) working correlation matrices. The GEE estimates were obtained using PROC GENMOD in SAS version 6.12. The standard errors of the regression parameters were computed using the model-robust, sandwich-type estimator; see (5.10) in Chaganty (1997).

Table 1
Regression analysis of a dental study data using GEE and C-QLS methods with AR(1) and unstructured working correlation matrices

Parameter	GEE				C-QLS			
	AR(1)		UNSTR		AR(1)		UNSTR	
	Est.	Std.	Est.	Std.	Est.	Std.	Est.	Std.
β_g	17.3213	0.7780	17.3973	0.7244	17.3215	0.7776	17.4018	0.6972
β_b	16.5946	1.2788	16.3236	1.1701	16.5931	1.2781	16.0523	1.1288
γ_g	0.4838	0.0629	0.4781	0.0639	0.4837	0.0629	0.4770	0.0632
γ_b	0.7965	0.1050	0.7881	0.0983	0.7695	0.1049	0.8122	0.0939
ϕ	4.9107		4.9058		4.9106		4.9076	

Table 2
Estimates of the working correlation matrices. C-QLS (GEE) estimates are above (below) the diagonal

AR(1)				UNSTR			
—	0.6099	0.3720	0.2269	—	0.5284	0.6609	0.5084
0.6135	—	0.6099	0.3720	0.5010	—	0.5551	0.7034
0.3764	0.6135	—	0.6099	0.7363	0.5553	—	0.7269
0.2309	0.3764	0.6135	—	0.5149	0.6208	0.7788	—

Table 3
Regression analysis of audiology data using C-QLS

CORR	$\hat{\beta}_0$ (SE)	$\hat{\beta}_1$ (SE)	$\hat{\beta}_2$ (SE)
MARK	16.99 (3.35)	2.31 (0.32)	−0.05 (0.01)
GMARK	18.67 (3.60)	1.92 (0.31)	−0.03 (0.01)

Estimates of the working correlation matrices are displayed in Table 2. It is clear from Tables 1 and 2 that the estimates obtained by GEE and C-QLS are in reasonable agreement, but with one exception. The standard errors of the regression parameter estimates obtained by C-QLS are smaller than, and hence preferable to, those obtained by GEE.

5.2. *Analysis of unbalanced and unequally spaced data*

In this section we apply the Markov and generalized Markov structures using the C-QLS approach to estimation of the parameters. As discussed in Shults and Chaganty (1998), the GEE method often yields infeasible correlation parameter estimates for the Markov structure. It is also difficult to apply GEE using the generalized Markov model because moment estimates for its parameters are not easy to obtain. Both these correlation structures were not implemented in the SAS, version 6.12, GEE procedure PROC GENMOD.

The data we examine (see Table 3 of Núñez-Anton and Woodworth, 1994) were also analyzed by Shults and Chaganty (1998). They comprise measurements collected during a study to compare two cochlear prostheses implanted in a group of postlingually deafened adults. The study outcome is the percentage of sentences recognized on a sentence recognition test that was administered at 1, 9, 18, and 30 months post implantation. Because not all subjects completed all four sentence recognition tests, the data are unbalanced and unequally spaced in time.

Our final regression model for the marginal mean of the outcome variable (μ_{ij}) agrees with the final model fit by Núñez-Anton and Woodworth (1994)

$$\mu_{ij} = \beta_0 + \beta_1 t_{i1} + \beta_2 t_{i2}, \tag{5.2}$$

where t_{i1} is the month of measurement and $t_{i2} = t_{i1}^2$. Table 3 contains estimates and standard errors for the regression parameters that were computed using C-QLS and the

Table 4
C-QLS estimates of the correlation between successive measurements

CORR	1 and 9 mo.	9 and 18 mo.	18 and 30 mo.
MARK	0.9168	0.9069	0.8778
GMARK	0.8442	0.9433	0.9564

Markov (MARK) and generalized Markov (GMARK) working correlation models. The robust standard errors were obtained using (5.10) in Chaganty (1997). As can be seen in Table 3, fitting the generalized Markov structure yields an estimated constant that is slightly greater in value and coefficients associated with the timings of measurements that are slightly smaller in value than the corresponding coefficients obtained by fitting the Markov model.

Most interesting however, is the difference between the correlation estimates that are obtained by fitting the generalized Markov as opposed to the simpler Markov model. The generalized Markov model allows us to model the intra-subject correlations in a manner consistent with the findings of Gantz et al. (1988), who observed that there is a 'definite learning curve involved with the use of cochlear implants'. This implies that the correlation between two measurements will be greater if the measurements are collected on a subject later during the study, rather than earlier. Table 4 contains the estimated correlation between two successive measurements on a subject for the Markov (MARK) and the generalized Markov (GMARK) correlation structures. (These estimates are based on $\hat{\rho}_g = 0.9892$ for the Markov structure and $(\hat{\tilde{\eta}}, \hat{\tilde{\lambda}}) = (0.9305, 0.0612)$ for the generalized Markov structure.) As is shown in Table 4, the generalized Markov structure yields what we expect for the audiology data- increasing intra-subject correlations over time, whereas the Markov structure forces a decrease in the correlation between the successive measures on each subject.

While modelling intra-subject correlation is not our primary goal, fitting the correlation structure that is most reasonable for our data analysis situation should yield the best results in terms of analysis of our main parameter of interest. In any case, it would be contrary to the tradition of statistical modelling to assume that fitting the model that does not best approximate reality would yield optimum results for our data analysis problem. The ability to fit the generalized Markov structure, the structure appropriate when the outcome variable stabilizes over time, is thus an important advantage afforded by the C-QLS approach over GEE.

Acknowledgements

This material is based upon work supported in part by the US Army research office under grant number DAAH04-96-1-0070. The United States Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon.

Appendix

Proof of Theorem 4.2. Let $\theta=(\beta, \rho)'$, α and ϕ be fixed. Assume that the working correlation structure $R(\alpha)$ is AR(1). Suppose that $\hat{\theta}_a=(\hat{\beta}_{am}, \hat{\alpha}_{am})'$ is the solution of Eqs. (2.2) and (2.3). Let $\hat{\theta}_c=(\hat{\beta}_{am}, \hat{\rho}_{am})'$, where $\hat{\rho}_{am}$ is defined in Eq. (3.13). Assume that the conditions of Theorem 5.1 in Chaganty (1997) hold. Let $v_i(\theta)=Z_i'(\beta)R^{-1}(\rho)Z_i(\beta)$. Since $\sum_{i=1}^m \nabla v_i(\hat{\theta}_a)=0$, using Taylor series expansions we can write

$$\sum_{i=1}^m v_i(\hat{\theta}_c)-\sum_{i=1}^m v_i(\theta)=(\hat{\theta}_c-\theta)'\sum_{i=1}^m \nabla v_i(\theta)+(\hat{\theta}_c-\theta)'\frac{1}{2}\sum_{i=1}^m \nabla^2 v_i(\theta^*)(\hat{\theta}_c-\theta)$$

$$=(\hat{\theta}_c-\theta)'\sum_{i=1}^m \nabla^2 v_i(\theta^{**})(\theta-\hat{\theta}_a)$$

$$+(\hat{\theta}_c-\theta)'\frac{1}{2}\sum_{i=1}^m \nabla^2 v_i(\theta^*)(\hat{\theta}_c-\theta), \tag{A.1}$$

where θ^*, θ^{**} are points on the line joining $\hat{\theta}_c$ and θ, and the line between $\hat{\theta}_a$ and θ, respectively. From Eq. (A.1) we get

$$\frac{1}{m}\left[\sum_{i=1}^m v_i(\theta)-\sum_{i=1}^m v_i(\hat{\theta}_c)\right]=(\hat{\theta}_c-\theta)'\frac{1}{m}\sum_{i=1}^m \nabla^2 v_i(\theta^{**})(\hat{\theta}_a-\theta)$$

$$-(\hat{\theta}_c-\theta)'\frac{1}{2m}\sum_{i=1}^m \nabla^2 v_i(\theta^*)(\hat{\theta}_c-\theta). \tag{A.2}$$

Since $(\hat{\theta}_c-\theta)\to 0$ and $(\hat{\theta}_a-\theta)$, $(1/m)\sum_{i=1}^m \nabla^2 v_i(\theta^*)$, $(1/m)\sum_{i=1}^m \nabla^2 v_i(\theta^{**})$ are bounded in probability, from Eq. (A.2) we get

$$\frac{1}{m}\left[\sum_{i=1}^m v_i(\theta)-\sum_{i=1}^m v_i(\hat{\theta}_c)\right]\to 0 \tag{A.3}$$

in probability as $m\to\infty$. By the weak law of large numbers we also have

$$\frac{1}{m}\sum_{i=1}^m v_i(\theta)\to \phi\ \mathrm{tr}(R^{-1}(\rho)\bar{R}), \tag{A.4}$$

in probability as $m\to\infty$. It is easy to verify that $\mathrm{tr}(R^{-1}(\rho)\bar{R})=n$ when $\bar{R}$ is any one of the following structures: (i) equicorrelated, (ii) AR(1), and (iii) tridiagonal. We can now see that Eq. (4.2) follows from Eqs. (A.3) and (A.4). When $\rho=0$ and the true correlation is the identity matrix, we can verify that $(\hat{\theta}_c-\theta)\to 0$ in probability as $m\to\infty$. Thus (A.3) also holds in this case. This completes the proof of the theorem. □

References

Chaganty, N.R., 1997. An alternative approach to the analysis of longitudinal data via generalized estimating equations. J. Statist. Plann. Inference 63, 39–54.

Gantz, B.J. et al., 1988. Evaluation of five different cochlear implant designs: audiologic assessment and predictors of performance. Laryngoscope 98, 1100–1106.

Jennrich, R.I., Schluchter, M.D., 1986. Unbalanced repeated-measures models with structured covariance matrices. Biometrics 42, 805–820.

Liang, K.Y., Zeger, S.L., 1986. Longitudinal data analysis using generalized linear models. Biometrika 73, 13–22.

Núñez-Anton, V., Woodworth, G.G., 1994. Analysis of longitudinal data with unequally spaced observations and time-dependent correlated errors. Biometrics 50, 445–456.

Olkin, I., Pratt, J.W., 1958. A multivariate Tchebycheff inequality. Ann. Math. Statist. 29, 226–234.

Potthoff, R.F., Roy, S.N., 1964. A generalized multivariate analysis of variance model useful especially for growth curve problems. Biometrika 51, 665–680.

Shults, J., Chaganty, N.R., 1998. Analysis of serially correlated data using quasi-least squares. Biometrics 54, 1622–1630.

Whittle, P., 1958. A multivariate generalization of Tchebichev's inequality. Quart. J. Math. Oxford Ser. 9, 232–240.

Loss in Efficiency Due to Misspecification of the Correlation Structure in GEE

Narasinga Rao Chaganty
Department of Mathematics and Statistics
Old Dominion University, Norfolk, VA, 23529.
U. S. A.

1. Preliminaries. In this paper we will need the following matrix version of the Cauchy-Schwartz inequality.

Theorem 1. Let B_i, D_i, $1 \le i \le m$, be matrices of order $t \times p$. Let Σ_i, $1 \le i \le m$ be positive definite matrices of order t. Then

$$\left(\sum_{i=1}^{m} B_i' D_i\right)^{-1} \left(\sum_{i=1}^{m} B_i' \Sigma_i B_i\right) \left(\sum_{i=1}^{m} D_i' B_i\right)^{-1} - \left(\sum_{i=1}^{m} D_i' \Sigma_i^{-1} D_i\right)^{-1} \qquad (1)$$

is nonnegative definite, assuming that the inverses in (1) exist.

2. Optimal Estimating function. The longitudinal data analysis problem considered by Liang & Zeger (1986) can be described briefly as follows: Let $\{Y_i, 1 \le i \le m\}$ be independent vectors such that $E(Y_i) = \mu_i(\beta)$ and covariance matrix $\Sigma_i = A_i^{1/2}(\beta) \overline{R} A_i^{1/2}(\beta)$, where $\overline{R}$ is the true correlation between the components of the vector Y_i assumed to be the same for all $1 \le i \le m$. The mean vector $\mu_i(\beta)$ and the diagonal variance matrix $A_i(\beta)$ are assumed to be known functions of $\beta_{p\times 1}$. The problem is to estimate the unknown regression parameter β. Following the ideas contained in Godambe (1960) and Godambe & Kale (1991), let us consider the class of unbiased estimating equations

$$\mathcal{G} = \left\{ \sum_{i=1}^{m} B_i' (Y_i - \mu_i(\beta)) = 0 \right\} \qquad (2)$$

where B_i, $1 \le i \le m$, are $t \times p$ matrices. Let $D_i(\beta) = \partial \mu_i / \partial \beta'$ be the matrix of partial derivatives of order $t \times p$. For each estimating function $g \in \mathcal{G}$, let

$$M_g = \left(\sum_{i=1}^{m} B_i' D_i\right)^{-1} \left(\sum_{i=1}^{m} B_i' \Sigma_i B_i\right) \left(\sum_{i=1}^{m} D_i' B_i\right)^{-1}. \qquad (3)$$

The matrix M_g in (3) is the covariance matrix of the standardized version, $g_s = (E(\partial g/\partial\beta))^{-1} g$, of the estimating function g (see Godambe & Kale (1991), p. 14). Note that when $B_i = \Sigma_i^{-1} D_i$, the matrix $M_g = M_{g^*}$, where $M_{g^*} = \left(\sum_{i=1}^m D_i' \Sigma_i^{-1} D_i\right)^{-1}$. By Theorem 1 we have that $M_g - M_{g^*}$ is nonnegative definite for all $g \in \mathcal{G}$. Therefore, if $\overline{R}$ is known the optimal estimating function for β in the class $\mathcal{G}$ is given by $g^* = \sum_{i=1}^m D_i' \Sigma_i^{-1} (Y_i - \mu_i(\beta))$. Since in practice the true correlation $\overline{R}$ is unknown, Liang & Zeger (1986) have suggested that we replace $\overline{R}$ by a working correlation structure $R(\alpha)$, which is assumed to be a function of α. The estimate of β is obtained by solving the generalized estimating equation (GEE), $g_w = \sum_{i=1}^m D_i' \Sigma_{wi}^{-1} (Y_i - \mu_i(\beta)) = 0$, where $\Sigma_{wi} = A_i^{1/2}(\beta)\, R(\alpha)\, A_i^{1/2}(\beta)$.

3. Loss in Efficiency. Since g^* is the optimal estimating function, a measure of efficiency of the estimating function g_w is given by $\text{eff}_d(g_w) = |M_{g^*}|/|M_{g_w}|$, if we use determinant optimality criteria or $\text{eff}_t(g_w) = \text{tr}(M_{g^*})/\text{tr}(M_{g_w})$, if we prefer the trace (tr) optimality criterion. The loss in efficiency due to misspecification is given by $L_d(g_w) = (1 - \text{eff}_d(g_w))$ or $L_t(g_w) = (1 - \text{eff}_t(g_w))$. The quantities $L_d(g_w)$ and $L_t(g_w)$ can be estimated by replacing β, α and $\overline{R}$ by the GEE estimates given in Liang & Zeger (1986). We suggest that, in the analysis of longitudinal data one should try several working correlation structures and choose the structure that yields the smallest value for $L_d(g_w)$ or $L_t(g_w)$.

BIBLIOGRAPHY

Liang, K-Y. & Zeger, S. L. (1986). Longitudinal data analysis using generalised linear models. *Biometrika* **73**, 13-22.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Stat.*, **31**, 1208-12.

Godambe, V. P. & Kale, B. K. (1991). Estimating functions: an overview. In *Estimating Functions*, (V. P. Godambe, ed.) 3-20. Oxford: Oxford University Press.

Summary: The concept of a working correlation structure was introduced by Liang & Zeger (1986) to analyze a longitudinal data analysis problem via generalized estimating equations (GEE). In this paper we first derive the optimal estimating function for the longitudinal data analysis problem using the criterion given in Godambe & Kale (1991). We then propose a method of estimating the loss in efficiency if the working correlation is not the true correlation structure, that is, if it is misspecified. Our method also serves as a tool for choosing a working correlation structure that maximizes efficiency in finite samples.

Résume. Liang et Zeger (1986) ont présenté le concepte d'une structure de corrélation afin d'analyser le problème d'analyse de données longitudinales par des équations généralisées d'estimation (EGE). Dans cette étude nous dérivons d'abord la fonction optimale de calcul approximatif pour le problème d'analyse de données longitudinales en utilisant les critères de Godambe et Kale (1991). Nous proposons ensuite une méthode pour calculer la perte d'efficacité si la corrélation proposée n'est pas l'exacte structure de corrélation, c'est-á-dire, si elle est mal spéficiée. Notre méthode sert également d'outil pour choisir une structure de corrélation qui porte au maximum l'efficacité des échantillons finis.

Analysis of the growth curve model using quasi-least squares

N. Rao Chaganty
Department of Mathematics and Statistics
Old Dominion University
Norfolk, VA 23529-0077, USA
email: nrc@math.odu.edu

1. Summary

The growth curve model of Potthoff and Roy (1964) has been studied extensively by numerous authors via the maximum likelihood method, under the assumption of normality for the outcome variable. In this paper we apply the method of quasi-least squares developed in Chaganty (1997) and Chaganty and Shults (1999) for estimating and testing a hypothesis concerning the parameters in the growth curve model when the normality assumption is not satisfied. Large sample properties of the estimates and the test statistic are also presented.

2. Quasi-least squares estimates

The growth curve model is defined as $\mathbf{Y}_{p\times n} = \mathbf{B}_{p\times m}\ \mathbf{\Delta}_{m\times r}\ \mathbf{A}_{r\times n} + \mathbf{E}_{p\times n}$ where $\mathbf{Y}$ is the response matrix consisting of p repeated measurements taking on n individuals, $\mathbf{\Delta}$ is an unknown parameter matrix, $\mathbf{A}$ and $\mathbf{B}$ are known within-subject and between-subject design matrices of ranks m and t respectively. We assume that the error matrix $\mathbf{E}$ has zero mean and $cov\,(vec\,(\mathbf{E})) = \sigma^2\,\mathbf{I}_n \otimes R(\alpha)$, where $vec\,(\mathbf{E})$ is the $p\,n \times 1$ column vector formed by stacking the columns of the matrix $\mathbf{E}$. Here $\mathbf{I}_n$ is the identity matrix and $R(\alpha)$ is a correlation matrix that is a function of the unknown parameter α. Several authors have studied this model under the assumption that $\mathbf{E}$ has a matrix normal distribution. We do not make any such assumption in this paper. The method of quasi-least squares developed in Chaganty (1997) and Chaganty and Shults (1999) is based on minimizing the objective function

$$Q(\mathbf{\Delta},\,\alpha) = \text{tr}\,\left((\mathbf{Y} - \mathbf{B}\,\mathbf{\Delta}\,\mathbf{A})'\,R^{-1}(\alpha)\,(\mathbf{Y} - \mathbf{B}\,\mathbf{\Delta}\,\mathbf{A})\right). \tag{1.1}$$

Equating to zero the partial derivatives of (1.1) with respect to $\mathbf{\Delta}$ and α, we get

$$\mathbf{\Delta} = (\mathbf{B}'\,R^{-1}(\alpha)\,\mathbf{B})^{-1}\,\mathbf{B}'\,R^{-1}(\alpha)\,\mathbf{Y}\,\mathbf{A}'(\mathbf{A}\,\mathbf{A}')^{-1} \tag{1.2}$$

and

$$\text{tr}\left(\frac{\partial}{\partial\,\alpha}\,R^{-1}(\alpha)\,\mathbf{U}(\mathbf{\Delta})\right) = 0 \tag{1.3}$$

where $\mathbf{U}(\mathbf{\Delta}) = (\mathbf{Y} - \mathbf{B}\,\mathbf{\Delta}\,\mathbf{A})\,(\mathbf{Y} - \mathbf{B}\,\mathbf{\Delta}\,\mathbf{A})'$. Let $(\widetilde{\mathbf{\Delta}},\ \tilde{\alpha})$ be the solution of the equations (1.2) and (1.3). We can show that $\tilde{\alpha}$ is asymptotically biased as $n \to \infty$, since the estimating equation (1.3) is not unbiased. However, the solution $\hat{\alpha}$ of the estimating equation

$$\text{tr}\left(\frac{\partial}{\partial\,\alpha}\,R^{-1}(\alpha)\Big|_{\alpha=\tilde{\alpha}}\,R(\alpha)\right) = 0 \tag{1.4}$$

is a consistent estimate of α. We shall call $\hat{\alpha}$ the quasi-least squares estimate of α. If the correlation matrix $R(\alpha)$ has the AR(1) structure, we can verify that $\hat{\alpha} = 2\,\tilde{\alpha}/(1+\tilde{\alpha}^2)$. The quasi-least squares estimates of $\mathbf{\Delta}$ and σ^2 are given by

$$\widehat{\mathbf{\Delta}} = (\mathbf{B}'\,\hat{R}^{-1}\,\mathbf{B})^{-1}\,\mathbf{B}'\,\hat{R}^{-1}\,\mathbf{Y}\,\mathbf{A}'\,(\mathbf{A}\,\mathbf{A}')^{-1} \text{ and } \hat{\sigma}^2 = \text{tr}\,\left(\hat{R}^{-1}\,\widehat{\mathbf{U}}\right)\Big)/pn \tag{1.5}$$

where $\widehat{\mathbf{U}} = \mathbf{U}(\widehat{\mathbf{\Delta}})$ and $\hat{R} = R(\hat{\alpha})$. The large sample properties of the quasi-least squares estimates are established in the following theorem:

THEOREM 1. *Fix $\mathbf{\Delta}$, α and σ^2. Let $\widehat{\mathbf{\Delta}}$, $\hat{\alpha}$ and $\hat{\sigma}^2$ be the quasi-least squares estimates of $\mathbf{\Delta}$, α and σ^2 respectively. Assume that the matrix $\mathbf{W} = \mathbf{A}'(\mathbf{A}\,\mathbf{A}')^{-1/2} = (w_{ij})$ is such that $\max_{ij}\, w_{ij}^2$ converges to zero as $n \to \infty$. Then as $n \to \infty$,*

(a) *vec $\left((\widehat{\mathbf{\Delta}} - \mathbf{\Delta})\,(\mathbf{A}\,\mathbf{A}')^{1/2}\right)$ converges in distribution to a multivariate normal distribution with mean zero and covariance matrix $\sigma^2\,\mathbf{I}_r \otimes (\mathbf{B}'\,R^{-1}(\alpha)\,\mathbf{B})^{-1}$.*

(b) *$\hat{\alpha} \to \alpha$ and $\hat{\sigma}^2$ converges to σ^2 in probability.*

3. Test of Hypothesis

We now consider the problem of testing $H_0 : D\,\mathbf{\Delta}\,E = N \quad vs \quad H_a : D\,\mathbf{\Delta}\,E \neq N$, where $D_{d\times m}$ and $E_{r\times e}$ are known matrices of ranks d and e respectively. In most situations the matrix N is the null matrix. Let $\overline{\mathbf{\Delta}}$ be the estimate obtained minimizing $Q(\mathbf{\Delta}, \hat{\alpha})$ with respect to $\mathbf{\Delta}$, subject to the restriction $D\,\mathbf{\Delta}\,E = N$. It is easy to verify that

$$\overline{\mathbf{\Delta}} = \widehat{\mathbf{\Delta}} + (\mathbf{B}'\,\hat{R}^{-1}\,\mathbf{B})^{-1}\,D'\,\overline{\lambda}\,E'(\mathbf{A}\,\mathbf{A}')^{-1} \tag{1.6}$$

where $\overline{\lambda} = (D\,(\mathbf{B}'\hat{R}^{-1}\,\mathbf{B})^{-1}\,D')^{-1}\,(N - D\widehat{\mathbf{\Delta}}\,E)\,(E'\,(\mathbf{A}\,\mathbf{A}')^{-1}\,E)^{-1}$. We propose the test statistic

$$T = \text{tr}\,\left((\mathbf{B}(\mathbf{B}'\,\hat{R}^{-1}\,\mathbf{B})^{-1}\,\mathbf{B}')^{-1}\,\mathbf{B}\,(\overline{\mathbf{\Delta}} - \widehat{\mathbf{\Delta}})\mathbf{A}\,\mathbf{A}'(\overline{\mathbf{\Delta}} - \widehat{\mathbf{\Delta}})'\,\mathbf{B}'\right)/\hat{\sigma}^2. \tag{1.7}$$

Large values of T are considered to be significant and the theorem below is useful for determining the critical values.

THEOREM 2. *Assume that the conditions of Theorem 1 hold. Then as $n \to \infty$, under the null hypothesis $H_0 : D\,\mathbf{\Delta}\,E = N$, the distribution of T converges to a central χ^2 with pr degrees of freedom.*

To test the hypothesis H_0, we could also use other multivariate tests based on the eigenvalues of the matrix in (1.7) instead of the trace.

4. REFERENCES

Chaganty, N. R. (1997). An alternative approach to the analysis of longitudinal data via generalized estimating equations. J. Statist. Plann. Inference 63, 39-54.

Chaganty, N. R. and Shults, J. (1999). On eliminating the asymptotic bias in the quasi-least squares estimate of the correlation parameter. J. Statist. Plann. Inference 76, 145-161.

Potthoff, R. F. and Roy, S. N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. Biometrika 51, 313-326.

Analysis of Multivariate Longitudinal Data Using Quasi-Least Squares

By

N. Rao Chaganty and Dayanand N. Naik
Department of Mathematics and Statistics
Old Dominion University
Norfolk, VA 23529.

Abstract

In this paper we consider the analysis of multivariate longitudinal data assuming a scale multiple of Kronecker product correlation structure for the covariance matrix of the observations on each individual. The method used for the estimation of the parameters is the quasi-least squares method introduced by Chaganty (1997, *J. Statist. Plann. Inference* 63, 39-54), and further developed by Shults and Chaganty (1998, *Biometrics* 54, 1622-1630) and Chaganty and Shults (1999, *J. Statist. Plann. Inference* 76, 145-161). We show that the estimating equations for the correlation parameters in the quasi-least squares method are optimal unbiased estimating equations if the data is from a normal population. An algorithm for computing the estimates is provided and implemented on two real life data sets. The asymptotic joint distribution of the estimators of the regression and correlation parameters is derived and used for testing a linear hypothesis on the regression parameters.

Key words: Longitudinal data, Kronecker product, Quasi-least squares, GEE, AR(1).
AMS 1991 Subject classifications: Primary 62H12, 62H15, 62J12; Secondary 62F10, 62F12.

1 Introduction

In many practical situations, observations on n experimental units (or subjects) are made on a set of p response variables (or characteristics) at t occasions. Thus on each experimental unit we have a $p \times t$ matrix of observations. These data are termed as *multivariate repeated measures* data, or *multivariate longitudinal* data, or *multi-response growth curve* data. Few examples of the experiments that yield multivariate longitudinal data are given below.

- In an experiment to study the effect of iron with vitamin C supplement, n subjects may be classified into one of the three groups: Group 1 receiving 15 mg elemental iron with 25 mg vitamin C three times a day, Group 2 receiving 15 mg iron with 50 mg vitamin C three times a day and Group 3 receiving simply 15 mg iron three times a day. Measurements, using the blood sample, may be collected on the variables: serum iron, ferritin, transferrin saturation, hemoglobin, hematocrit, and total iron binding capacity (TIBC). Observations on these six variables ($p = 6$) may be made at each of the three time periods ($t = 3$).

- In an experiment where a new drug for AIDS is being tested, on each of the n subjects data on three variables ($p = 3$) (TMHR scores, Karofsky scores, and T-4 cell counts) at three time periods ($t = 3$) during the study (at the beginning, after 90 days of treatment, and after 180 days of treatment) are collected. These data will be analyzed in Section 5 of this paper.

- An experiment in dental study concerns with the relative effectiveness of two orthopedic adjustments of the mandible. Nine subjects are assigned to each of the two orthopedic treatment groups known as activator treatments. The measurements are made on three characteristics ($p = 3$) to assess the changes in the vertical position of the mandible at three time points ($t = 3$) of activator treatment. We will also analyze these data in Section 5.

Suppose we have n subjects (possibly randomly assigned to g groups) on which the measurements are made on p response variables at t occasions. Let y_{ijk} be the observation on the jth response variable taken at the kth time period or occasion corresponding to the ith individual. Here $1 \leq i \leq n$, $1 \leq j \leq p$, $1 \leq k \leq t$. Also, associated with each of the n subjects, suppose we have measurements x_{ijkl} taken on q covariates ($1 \leq l \leq q$). The covariates could be categorical, and they may or may not change with time. Let

$$x_{il} = (x_{i11l}, ..., x_{ip1l}, x_{i12l}, ..., x_{ip2l}, ..., x_{i1tl}, ..., x_{iptl})'$$

be the vector of observations on the lth covariate taken on the ith subject. Let $X_i = [x_{i1} : \ldots : x_{iq}]_{pt \times q}$ be the matrix of measurements taken on the q covariates associated with each response variable at the t occasions on the ith individual and

$$Y_i = \begin{bmatrix} y_{i11} & \cdots & y_{i1t} \\ \vdots & \ddots & \vdots \\ y_{ip1} & \cdots & y_{ipt} \end{bmatrix}_{p \times t}$$

be the matrix of measurements taken on the response variables on the ith individual. Suppose that the expected value and the covariance matrix of $y_i = vec(Y_i)$ are $E(y_i) = \mu_i(\beta) = X_i\beta$ and $Cov(y_i) = \Omega$ respectively. Analysis of these data is complicated by the existence of correlation among the measurements on p different variables together with the correlation among measurements taken at t different occasions. The form of the covariates, that is, whether they are subject specific, time varying, or varying with the response variables may further complicate the analysis. However, assuming that the data on each individual come from a pt dimensional multivariate normal distribution with dispersion matrix Ω, the maximum likelihood estimator of β can be obtained as a function of Ω and inference can be performed using the standard asymptotic theory of maximum likelihood estimators. Some general discussion of this, using a mixed model approach, in an applied point of view can be found in Khattree and Naik (1999).

If the data do not come from a multivariate normal distribution or if the response variables on which data are collected are not continuous then the standard methods do not readily apply. Recently, Chaganty (1997) introduced a new method called "quasi-least squares" for analyzing longitudinal data. This method is an alternative to the GEE method of Liang and Zeger (1986) and its various variations. Quasi-least squares method was developed to overcome some of the pitfalls of the GEE method (Crowder, 1995). Unlike the GEE method which can yield non-feasible and inconsistent estimates for the correlation parameters, Chaganty (1997) and Chaganty and Shults (1999) have shown that the "quasi-least squares method" always yields feasible and consistent estimates for the correlation parameters. Quasi-least squares method has been successfully utilized in various practical problems involving unbalanced and unequally spaced data. See Shults and Chaganty (1998) and Chaganty and Shults (1999).

It has been observed by Boik (1991) and Naik and Rao (1997) that assuming a Kronecker product structured covariance, that is, $\Omega = \Omega_1 \otimes \Omega_2$, where Ω_1 and Ω_2 respectively are $t \times t$ and $p \times p$ positive definite matrices, has many advantageous in analyzing multivariate repeated measures data. Further the linear model with this covariance structure reduces to the well known Zellner's Seemingly Unrelated Regression (SUR) model when $\Omega_2 = I$. Hence in this article we will consider Kronecker product covariance structure for the dispersion matrix of y_i.

The main focus of this paper is to implement the quasi-least squares method for analyzing multivariate longitudinal data assuming a scale multiple of Kronecker product correlation structure for the covariance matrix. The organization of the paper is as follows. In Section 2,

we will describe the quasi-least squares method as applied to the present situation. We also present a discussion of the optimality of the estimating equations and an iterative algorithm for the computation of the estimates. In Section 3, we will derive closed form solutions for the estimates of the correlation parameters for some popular correlation structures. In Section 4, we will derive the joint asymptotic distribution of the quasi-least squares regression and correlation parameter estimates. We also present a test statistic for testing linear hypothesis concerning the regression parameter β and derive its asymptotic distribution. We will present the analysis of two data sets in Section 5 and finally end with some concluding remarks.

2 The method of quasi-least squares

For analyzing multivariate repeated measures data that are continuous non-normal or categorical we adopt the quasi-least squares method, described in Chaganty (1997), and the bias corrected version of the correlation parameter in Chaganty and Shults (1999). To put the problem in a slightly general frame work we assume that

$$E(y_i) = \mu_i(\beta) = g(X_i\beta), \tag{2.1}$$

where as before X_i is the $pt \times q$ design matrix, β is a $q \times 1$ vector of unknown parameters and the inverse of g is a known link function. Further assume that the covariance matrix of y_i is

$$\Omega = \phi\, A_i^{1/2}(\beta)\, (R_T(\alpha) \otimes R_P(\gamma))\, A_i^{1/2}(\beta) = \phi\, \Sigma_i(\theta) \quad \text{(say)} \tag{2.2}$$

where $\theta = (\beta, \alpha, \gamma)'$ and $R_T(\alpha)$ and $R_P(\gamma)$ respectively are correlation matrices of order $t \times t$ and $p \times p$, which are functions of the vectors α and γ respectively. The correlation matrix $R_T(\alpha)$ represents the correlation among the t repeated measurements over time, whereas, $R_P(\gamma)$ represents the correlation among the p response variables. The $pt \times pt$ diagonal matrix $A_i^{1/2}(\beta)$ contains the standard deviations and ϕ is an overdispersion or a scale parameter. The mean-covariance model (2.1)-(2.2) encompasses several discrete and continuous models. While the main parameter of interest is β, the parameters α, γ and ϕ are nuisance parameters.

2.1 Estimating equations

Here we describe the method of quasi-least squares. This is a two stage procedure. In the first stage we minimize with respect to β, α and γ the quadratic form

$$\sum_{i=1}^{n} (y_i - \mu_i(\beta))' A_i^{-1/2}(\beta) (R_T^{-1}(\alpha) \otimes R_P^{-1}(\gamma)) A_i^{-1/2}(\beta) (y_i - \mu_i(\beta))$$

$$= \sum_{i=1}^{n} z_i(\beta)' (R_T^{-1}(\alpha) \otimes R_P^{-1}(\gamma)) z_i(\beta) \qquad (2.3)$$

where $z_i(\beta) = A_i^{-1/2}(\beta)(y_i - \mu_i(\beta))$ with $(A_i^{1/2}(\beta))^{-1} = A_i^{-1/2}(\beta)$. Note that $z_i(\beta) = vec\,(Z_i(\beta))$ where

$$Z_i(\beta) = \begin{bmatrix} z_{i11} & \cdots & z_{i1t} \\ \vdots & \ddots & \vdots \\ z_{ip1} & \cdots & z_{ipt} \end{bmatrix}_{p \times t}.$$

Since

$$\begin{aligned} \mathrm{tr}\left(R_T^{-1}(\alpha)\, Z_i'(\beta)\, R_P^{-1}(\gamma)\, Z_i(\beta)\right) &= vec\,(Z_i(\beta))' (R_T^{-1}(\alpha) \otimes R_P^{-1}(\gamma))\, vec\,(Z_i(\beta)) \\ &= z_i'(\beta)\, (R_T^{-1}(\alpha) \otimes R_P^{-1}(\gamma))\, z_i(\beta) \end{aligned}$$

we can rewrite the quadratic form (2.3) as

$$\mathrm{tr}(R_T^{-1}(\alpha) \textstyle\sum_{i=1}^{n} Z_i'(\beta)\, R_P^{-1}(\gamma)\, Z_i(\beta)) = n\,p\,\mathrm{tr}(R_T^{-1}(\alpha)\, U(\beta, \gamma)) \qquad (2.4)$$

and also as

$$\mathrm{tr}(R_P^{-1}(\gamma) \textstyle\sum_{i=1}^{n} Z_i(\beta)\, R_T^{-1}(\alpha)\, Z_i'(\beta)) = n\,t\,\mathrm{tr}(R_P^{-1}(\gamma)\, V(\beta, \alpha)) \qquad (2.5)$$

where the matrix $U(\beta, \gamma) = \dfrac{1}{n\,p} \displaystyle\sum_{i=1}^{n} Z_i'(\beta)\ R_P^{-1}(\gamma)\ Z_i(\beta)$ is of the order $t \times t$ and $V(\beta, \alpha) = \dfrac{1}{n\,t} \displaystyle\sum_{i=1}^{n} Z_i(\beta)\, R_T^{-1}(\alpha)\, Z_i'(\beta)$ is of order $p \times p$. Equating to zero the partial derivatives of (2.3), (2.4) and (2.5) with respect to β, α and γ respectively, we obtain the following three estimating equations:

$$\sum_{i=1}^{n} D_i'(\beta) A_i^{-1/2}(\beta) (R_T^{-1}(\alpha) \otimes R_P^{-1}(\gamma))\, z_i(\beta) = 0 \qquad (2.6)$$

$$\mathrm{tr}\left(\frac{\partial R_T^{-1}(\alpha)}{\partial \alpha} U(\beta, \gamma)\right) = 0 \qquad (2.7)$$

$$\mathrm{tr}\left(\frac{\partial R_P^{-1}(\gamma)}{\partial \gamma} V(\beta, \alpha)\right) = 0 \qquad (2.8)$$

where $D_i(\beta) = \partial \mu_i / \partial \beta'$. Let $\tilde{\theta} = (\tilde{\beta}, \tilde{\alpha}, \tilde{\gamma})'$ be the solution of the above three equations. The estimate $\tilde{\beta}$ is consistent but the estimates $\tilde{\alpha}$ and $\tilde{\gamma}$ are asymptotically biased (see Theorem 4.1).

The main reason being the estimating equation (2.6) is unbiased whereas the equations (2.7) and (2.8) are not unbiased. The second stage in the quasi-least squares method consists of solving the two equations

$$\text{tr}\left(\left.\frac{\partial R_T^{-1}(\alpha)}{\partial \alpha}\right|_{\alpha=\widetilde{\alpha}} R_T(\alpha)\right) = 0 \tag{2.9}$$

and

$$\text{tr}\left(\left.\frac{\partial R_P^{-1}(\gamma)}{\partial \gamma}\right|_{\gamma=\widetilde{\gamma}} R_P(\gamma)\right) = 0 \tag{2.10}$$

to obtain consistent estimates $\widehat{\alpha}$ and $\widehat{\gamma}$ of α and γ respectively. Let $\widehat{\beta} = \widetilde{\beta}$ (or the estimate obtained solving the equation (2.6) substituting $\widehat{\alpha}$ and $\widehat{\gamma}$ for α and γ respectively). We shall call the estimates $\widehat{\beta}$, $\widehat{\alpha}$ and $\widehat{\gamma}$ as the quasi-least squares estimates of β, α and γ respectively. Finally, a consistent estimate of ϕ is given by

$$\widehat{\phi} = \min\left(\widehat{\phi}_1, \widehat{\phi}_2\right) \tag{2.11}$$

where

$$\widehat{\phi}_1 = \frac{\sum_{i=1}^{n}\left(y_i - \mu_i(\widehat{\beta})\right)' \left(R_T(\widehat{\alpha}) \otimes R_P(\widehat{\gamma})\right)^{-1}\left(y_i - \mu_i(\widehat{\beta})\right)}{ntp}$$

and

$$\widehat{\phi}_2 = \frac{\sum_{i=1}^{n}\left(y_i - \mu_i(\widehat{\beta})\right)'\left(y_i - \mu_i(\widehat{\beta})\right)}{ntp}.$$

2.2 Optimality of the estimating equations

It is well known that, when α and γ are known, the function

$$g_1(\beta, \alpha, \gamma) = \sum_{i=1}^{n} D_i'(\beta) A_i^{-1/2}(\beta)\left(R_T^{-1}(\alpha) \otimes R_P^{-1}(\gamma)\right) z_i(\beta) \tag{2.12}$$

is the optimal unbiased estimating function for estimating β according to Godambe's criterion (see Godambe (1960), Heyde (1997, page 22)). Since $E(U(\beta, \gamma)) = \phi R_T(\alpha)$, the function

$$g_2(\beta, \alpha, \gamma, \phi) = \text{tr}\left(\frac{\partial R_T^{-1}(\alpha)}{\partial \alpha}(U(\beta, \gamma) - \phi R_T(\alpha))\right)$$

is clearly unbiased. Also, since $\frac{\partial R_T^{-1}(\alpha)}{\partial \alpha} = -R_T^{-1}(\alpha) \frac{\partial R_T(\alpha)}{\partial \alpha} R_T^{-1}(\alpha)$, using the properties of trace and Kronecker product (Rao and Rao, 1998, page 202) we can verify that

$$g_2(\beta, \alpha, \gamma, \phi) = -\left(\frac{\partial vec(R_T(\alpha))}{\partial \alpha}\right)' (R_T(\alpha) \otimes R_T(\alpha))^{-1} (vec(U(\beta, \gamma)) - \phi\, vec(R_T(\alpha))). \tag{2.13}$$

When β, γ and ϕ are known, the unbiased estimating function (2.13) is the optimal estimating function for estimating α *if* a constant multiple of Cov $(vec(U(\beta, \gamma)))$ is used in place of $(R_T(\alpha) \otimes R_T(\alpha))$. But Cov$(vec(U(\beta, \gamma)))$ depends in general on the fourth moments of the y_i's and we have no assumptions made concerning the fourth moments. However, note that if the y_i's are normal then Cov $(vec(U(\beta, \gamma)))$ is $2\,\phi\, (R_T(\alpha) \otimes R_T(\alpha))$. Thus the estimating function (2.13) is the optimal unbiased estimating function for the parameter α when the y_i's are normally distributed and the other parameters are known. And it will be close to the optimal unbiased estimating equation whenever a constant multiple of Cov $(vec(U(\beta, \gamma)))$ is approximately equal to $(R_T(\alpha) \otimes R_T(\alpha))$. Similarly, since $E(V(\beta, \alpha)) = \phi\, R_P(\gamma)$, the function

$$g_3(\beta, \alpha, \gamma, \phi) = \text{tr}\left(\frac{\partial R_P^{-1}(\gamma)}{\partial \gamma}(V(\beta, \alpha) - \phi\, R_P(\gamma))\right)$$

is unbiased. And if the y_i's are independent and normally distributed we can check that Cov $(vec(V(\beta, \alpha))) = 2\,\phi\, (R_P(\gamma) \otimes R_P(\gamma))$. Therefore, the function

$$g_3(\beta, \alpha, \gamma, \phi) = -\left(\frac{\partial vec(R_P(\gamma))}{\partial \gamma}\right)' (R_P(\gamma) \otimes R_P(\gamma))^{-1} (vec(V(\beta, \alpha)) - \phi\, vec(R_P(\gamma))) \tag{2.14}$$

is the optimal unbiased estimating function for estimating γ when β, α and ϕ are known, if the y_i's are normally distributed, and is close to being optimal if a constant multiple of Cov $(vec(V(\beta, \alpha)))$ is approximately equal to $(R_P(\gamma) \otimes R_P(\gamma))$. Now from (2.7), (2.8), (2.9) and (2.10) we can see that the quasi-least squares estimates satisfy $g_1(\widehat{\beta}, \widehat{\alpha}, \widehat{\gamma}) = 0$,

$$\text{tr}\left(\left.\frac{\partial R_T^{-1}(\alpha)}{\partial \alpha}\right|_{\alpha=\widetilde{\alpha}} (U(\widehat{\beta}, \widehat{\gamma}) - \phi\, R_T(\widehat{\alpha}))\right) = 0 \tag{2.15}$$

and

$$\text{tr}\left(\left.\frac{\partial R_P^{-1}(\gamma)}{\partial \gamma}\right|_{\gamma=\widetilde{\gamma}} (V(\widehat{\beta}, \widehat{\alpha}) - \phi\, R_P(\widehat{\gamma}))\right) = 0, \tag{2.16}$$

for all ϕ. In particular the equations (2.15) and (2.16) are satisfied when $\phi = \widehat{\phi}$. Thus the method of quasi-least squares provides a feasible solution to the unbiased estimating equations

$g_i = 0$, for $i = 1, 2, 3$. If the data is from a normal population then these three equations are also the optimal unbiased estimating equations according to Godambe's criterion. Regardless of normality, closeness of these estimating equations to optimality corresponds to closeness of (constant multiples of) Cov $(vec(U(\beta, \gamma)))$ and Cov $(vec(V(\beta, \alpha)))$ to $(R_T(\alpha) \otimes R_T(\alpha))$ and $(R_P(\gamma) \otimes R_P(\gamma))$, respectively.

2.3 Algorithm

In general a closed form solution does not exist for the estimating equations (2.6), (2.7) and (2.8). And we need to solve those equations using a recursive procedure like the Newton-Raphson method. An iterative algorithm for obtaining the first stage quasi-least squares estimates of β, α and γ could be described as follows:

Step 1: Start with a trial value β_0.

Step 2: Fix a trial value for γ_0 and compute $U_0 = U(\beta_0, \gamma_0)$.

Step 3: Get the estimate α_0 minimizing tr $(R_T^{-1}(\alpha)\, U_0)$ with respect to α.

Step 4: Compute $V_0 = V(\beta_0, \alpha_0)$.

Step 5: Get the estimate γ_1 minimizing tr$(R_P^{-1}(\gamma)\ V_0)$ with respect to γ.

Step 6: Repeat Steps 2 through 5 with $\gamma_0 = \gamma_1$, until convergence and obtain (γ_0, α_0).

Step 7: Compute the updated value

$$\beta_1 = \beta_0 + \left[\sum_{i=1}^{n} D_{i0}'\, \Sigma_{i0}^{-1}\, D_{i0}\right]^{-1} \left[\sum_{i=1}^{n} D_{i0}'\, \Sigma_{i0}^{-1}\, z_i(\beta_0)\right],$$

where $\Sigma_{i0} = \Sigma_i(\theta_0)$, $\theta_0 = (\beta_0, \alpha_0, \gamma_0)'$, $D_{i0} = D_i(\beta_0)$ and $D_i(\beta) = \partial\mu_i(\beta)/\partial\beta'$. Stop the iterative procedure if $\beta_1 \approx \beta_0$ and set $\widetilde{\beta} = \beta_0$, $\widetilde{\alpha} = \alpha_O$ and $\widetilde{\gamma} = \gamma_0$. Otherwise repeat Steps 2 through 6 with β_0 replaced by β_1.

We note that for most of the commonly used correlation structures the second stage in the quasi-least squares method does not require an iterative procedure, since the estimates can be obtained in a closed form as shown in the next section.

3 Correlation Structures

Here we consider several popular correlation structures, including the unstructured correlation for $R_T(\alpha)$ and illustrate the method of minimization of $\mathrm{tr}\,(R_T^{-1}(\alpha)\,U_0)$ needed in Step 3 of the algorithm. We will also obtain feasible, unique and often closed form solution to the equation (2.9) for these correlation structures. The structures assumed here for $R_T(\alpha)$ can be assumed for $R_P(\gamma)$ as well. And the form of the solutions $\tilde{\alpha}$ and $\hat{\alpha}$ obtained here for α can be used for obtaining $\tilde{\gamma}$ in the algorithm described in Section 2.3 and the solution $\hat{\gamma}$ for the equation (2.10) in Section 2.1.

3.1 Equicorrelated Correlation Structure

Suppose that the t repeated measurements are equicorrelated, that is, the correlation structure $R_T(\alpha)$ is of the form $R_T(\alpha) = (1-\alpha)\,I + \alpha\,J$, where I is the identity matrix and J is a matrix of ones. Since

$$R_T^{-1}(\alpha) = \frac{1}{(1-\alpha)}\,I - \frac{\alpha}{(1-\alpha)(1+(t-1)\alpha)}\,J$$

we have

$$\begin{aligned}\mathrm{tr}(R_T^{-1}(\alpha)\,U_0) &= \frac{1}{(1-\alpha)}\,\mathrm{tr}(U_0) - \frac{\alpha}{(1-\alpha)(1+(t-1)\alpha)}\,\mathrm{tr}(J\,U_0) \\ &= \frac{a}{(1-\alpha)} - \frac{\alpha\,b}{(1-\alpha)(1+(t-1)\alpha)}, \end{aligned} \tag{3.17}$$

where $a = \mathrm{tr}(U_0)$, $b = \mathrm{tr}(J\,U_0)$. Taking derivatives, we can check that the function (3.17) has a unique point of minimum in the interval $(-1/(t-1),\ 1)$, given by

$$\tilde{\alpha} = \frac{-a(t-1) + \sqrt{b\,(t-1)\,(at-b)}}{(t-1)(a\,(t-1)-b)}.$$

Also in this case there is a unique solution to the equation (2.9) in the interval $(-1/(t-1),\ 1)$ and it is given by

$$\hat{\alpha} = \kappa_1(\tilde{\alpha}) = \frac{\tilde{\alpha}^2\,(t-2) + 2\,\tilde{\alpha}}{[1+\tilde{\alpha}^2\,(t-1)]}. \tag{3.18}$$

We can verify that the function $\kappa_1(\cdot)$ is a continuous, one-to-one and onto function on the interval $(-1/(t-1),\ 1)$.

3.2 First Order Autoregressive (AR(1)) Correlation Structure

Consider the situation where the correlation between the t repeated measurements decreases with time. A commonly used correlation structure in this situation is the AR(1) structure, $R_T(\alpha) = [\alpha^{|i-j|}]$. Here

$$R_T^{-1}(\alpha) = \frac{1}{(1-\alpha^2)}[I - \alpha C_1 + \alpha^2 C_2],$$

where C_1 is a tridiagonal matrix with 0 on the diagonal and 1 on the lower and upper diagonals and $C_2 = \text{diag}(0, 1, ..., 1, 0)$. Therefore

$$\begin{aligned} \text{tr}(R_T^{-1}(\alpha)\, U_0) &= \frac{1}{(1-\alpha^2)}\,[\text{tr}(U_0) - \alpha\,\text{tr}(C_1\, U_0) + \alpha^2\,\text{tr}(C_2\, U_0)] \\ &= \frac{1}{(1-\alpha^2)}\,[a - \alpha\, c_1 + \alpha^2\, c_2], \end{aligned} \tag{3.19}$$

where $a = \text{tr}(U_0)$, $c_1 = \text{tr}(C_1\, U_0)$, and $c_2 = \text{tr}(C_2\, U_0)$. We can easily check that (3.19) has a unique point of minimum in the interval $(-1,\ 1)$ given by

$$\widetilde{\alpha} = \frac{(a + c_2) - \sqrt{(a + c_2)^2 - c_1^2}}{c_1}. \tag{3.20}$$

In this case the feasible solution to the equation (2.9) is

$$\widehat{\alpha} = \kappa_1(\widetilde{\alpha}) = \frac{2\,\widetilde{\alpha}}{(1 + \widetilde{\alpha}^2)}. \tag{3.21}$$

The function $\kappa_1(\cdot)$ is a continuous and one-to-one and onto function on the interval $(-1,\ 1)$. We will use the above estimate $\widehat{\alpha}$ in the examples discussed in Section 5.

3.3 Tri-Diagonal Structure

Let $R_T(\alpha)$ be a tri-diagonal matrix, that is, the diagonal elements of $R_T(\alpha)$ are one and all the elements above and immediately below the diagonal are equal to α and other elements are zero. Here $R_T^{-1}(\alpha)$ does not have a closed form but the matrix $R_T(\alpha)$ admits a spectral value decomposition

$$R_T(\alpha) = P\ \Lambda(\alpha)\ P'$$

where P, the matrix of orthogonal eigen vectors, does not depend on α. See Chaganty (1997), Example 4.2. Now

$$\text{tr}(R_T^{-1}(\alpha)U_0) \;=\; \text{tr}(\Lambda^{-1}(\alpha)\, P'\, U_0\, P)$$

$$= \sum_{k=1}^{t} \frac{u_k}{1 + 2\alpha \cos(\frac{k\pi}{t+1})} \tag{3.22}$$

where u_k is the k^{th} diagonal element of $P' U_0 P$. It is well known that $R_T(\alpha)$ is positive definite if and only if α falls in the interval (α_1, α_t), where

$$\alpha_j = \frac{-1}{2 \cos(\frac{j\pi}{t+1})}.$$

We can verify that (3.22) has a unique point of minimum $\tilde{\alpha}$ in the interval (α_1, α_t), which could be computed numerically. In this case the second stage estimate $\hat{\alpha}$ is in a closed form and is given by

$$\hat{\alpha} = \kappa_1(\tilde{\alpha}) = -\frac{1}{2} \frac{\sum_{k=1}^{t} b_k}{\sum_{k=1}^{t} b_k \cos(k\pi/(t+1))}$$

where

$$b_k = \frac{\cos(k\pi/(t+1))}{(1 + 2\tilde{\alpha} \cos(k\pi/(t+1)))}.$$

3.4 Unstructured Correlation matrix

Suppose that the correlation between the t repeated measurements $R_T(\alpha) = R_T$ is an unstructured positive definite correlation matrix. As shown in Chaganty (1997), the point of minimum $\tilde{R}_T$ in Step 3 of the algorithm described in Section 2.3, can be obtained recursively starting with any positive definite diagonal matrix Λ_0 and computing $\Lambda_k = \text{diag}(\Lambda_{k-1}^{1/2} U_0 \Lambda_{k-1}^{1/2})^{1/2}$ at the kth step and stop the recursive process as soon as $\Lambda_k \approx \Lambda_{k-1} = \tilde{\Lambda}$. The matrix $\tilde{R}_T = \tilde{\Lambda}^{-1/2} (\tilde{\Lambda}^{1/2} U_0 \tilde{\Lambda}^{1/2})^{1/2} \tilde{\Lambda}^{-1/2}$. The bias corrected correlation matrix is given by

$$\hat{R}_T = \begin{cases} \tilde{R}_T \Delta_T \tilde{R}_T & \text{if } \Delta_T > 0 \\ (\text{diag}(U_0))^{-1/2} U_0 (\text{diag}(U_0))^{-1/2} & \text{otherwise.} \end{cases} \tag{3.23}$$

where $\Delta_T = \text{diag}[(\tilde{R}_T \circ \tilde{R}_T)^{-1} e]$ where e is a vector of ones and $\circ$ denotes the Hadamard product (see Chaganty and Shults (1999)). Similarly, we can construct an estimate $\hat{R}_P$ for $R_P(\gamma) = R_P$, when it is an unknown unstructured correlation matrix. We will use the estimate $\hat{R}_P$ in the examples described in Section 5.

4 Large sample inference

In this section we will study the large sample properties of the quasi-least squares estimates. We show that the estimates are consistent and asymptotically normal. We also propose a test statistic for testing a hypothesis concerning the regression parameter and derive it's asymptotic distribution.

4.1 Asymptotic distribution

Here we will establish consistency and joint asymptotic normality of the quasi-least squares estimates. Let $\theta = (\beta, \alpha, \gamma)'$ be the vector consisting of the regression and the correlation parameters. Note that the first stage quasi-least squares estimate $\widetilde{\theta}$ is the solution of the equation $\sum_{i=1}^{n} h_i(\theta) = 0$, where $h_i(\theta) = (h_{1i}(\theta), h_{2i}(\theta), h_{3i}(\theta))'$ and

$$\begin{aligned} h_{1i}(\theta) &= D_i'(\beta) A_i^{-1/2}(\beta) \left(R_T^{-1}(\alpha) \otimes R_P^{-1}(\gamma)\right) z_i(\beta) \\ h_{2i}(\theta) &= \mathrm{tr}\left(\frac{\partial R_T^{-1}(\alpha)}{\partial \alpha} Z_i'(\beta) R_P^{-1}(\gamma) Z_i(\beta)\right) \\ h_{3i}(\theta) &= \mathrm{tr}\left(\frac{\partial R_P^{-1}(\gamma)}{\partial \gamma} Z_i(\beta) R_T^{-1}(\alpha) Z_i'(\beta)\right). \end{aligned}$$

The expected value of $h_i(\theta)$ does not depend on i and equals

$$\nu(\theta) = \left(0, \phi\,\mathrm{tr}\left(\frac{\partial R_T^{-1}(\alpha)}{\partial \alpha} R_T(\alpha)\right), \phi\,\mathrm{tr}\left(\frac{\partial R_P^{-1}(\gamma)}{\partial \gamma} R_P(\gamma)\right)\right)'. \tag{4.24}$$

Since $E(z_i(\beta)) = 0$, we can check that $I_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n} E\left(\frac{\partial h_i(\theta)}{\partial \theta'}\right)$ is of the form

$$I_n(\theta) = \begin{bmatrix} I_{n11}(\theta) & 0 & 0 \\ 0 & I_{n22}(\theta) & I_{n23}(\theta) \\ 0 & I_{n23}'(\theta) & I_{n33}(\theta) \end{bmatrix}. \tag{4.25}$$

In the above the three partitions are made according to the dimensions of the three vectors β, α and γ respectively. Similarly, we can partition $M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Cov}\,(h_i(\theta))$ as

$$M_n(\theta) = \begin{bmatrix} M_{n11}(\theta) & M_{n12}(\theta) & M_{n13}(\theta) \\ M_{n12}'(\theta) & M_{n22}(\theta) & M_{n23}(\theta) \\ M_{n13}'(\theta) & M_{n23}'(\theta) & M_{n33}(\theta) \end{bmatrix} \tag{4.26}$$

where $M_{njk} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Cov}(h_{ji}(\theta), h_{ki}(\theta))$. We can check that $M_{n11}(\theta) = \phi\, I_{n11}(\theta)$, where

$$I_{n11}(\theta) = \frac{1}{n}\sum_{i=1}^{n} D_i'(\beta)\,\Sigma_i^{-1}(\theta)\,D_i(\beta). \tag{4.27}$$

It is possible to express the other matrices M_{njk} and I_{njk} in (4.25) and (4.26) as functions of β, α, γ and ϕ explicitly. See Chaganty (1997, page 47) for some details concerning those formulas. The next theorem establishes the asymptotic normality of the first stage quasi-least squares estimates. Below we will use the acronym AN for asymptotically normal as in Serfling (1980, page 20).

THEOREM 4.1 *Let $\theta = (\beta, \alpha, \gamma)'$ be fixed. Let $\widetilde{\theta} = (\widetilde{\beta}, \widetilde{\alpha}, \widetilde{\gamma})'$ be the solution of the equation $\sum_{i=1}^{n} h_i(\theta) = 0$. Let $M_n(\theta) \to M(\theta)$ and $I_n(\theta) \to I(\theta)$ as $n \to \infty$. Assume that a central limit theorem holds for the summands $h_i(\theta)$ and they satisfy, as a function of θ, the regularity conditions needed for a Taylor series expansion to hold. Then*

$$\sqrt{n}\,(\widetilde{\theta} - \theta - [I(\theta)]^{-1}\,\nu(\theta)) \to N(0,\ [I(\theta)]^{-1}\,M(\theta)\,[I(\theta)]^{-1}) \tag{4.28}$$

as $n \to \infty$, where $\nu(\theta)$ is defined in (4.24).

Proof: Since $\sum_{i=1}^{n} h_i(\widetilde{\theta}) = 0$, using a Taylor series expansion and a standard argument we can verify that the asymptotic distribution of $(\widetilde{\theta} - \theta)$ is same as the asymptotic distribution of

$$\left[-\frac{1}{n}\sum_{i=1}^{n} E\left(\frac{\partial\, h_i(\theta)}{\partial\,\theta}\right)\right]^{-1}\left[\frac{1}{n}\sum_{i=1}^{n} h_i(\theta)\right] = [I_n(\theta)]^{-1}\left[\frac{1}{n}\sum_{i=1}^{n} h_i(\theta)\right]. \tag{4.29}$$

Note that $E(h_i(\theta)) = \nu(\theta)$ for all i. Since $M_n(\theta)$ converges to $M(\theta)$ and the summands $h_i(\theta)$ satisfy a central limit theorem, we conclude that

$$\left[\frac{1}{n}\sum_{i=1}^{n} h_i(\theta)\right] \quad is \quad AN\left(\nu(\theta),\ \frac{M(\theta)}{n}\right) \tag{4.30}$$

Since $I_n(\theta)$ converges to $I(\theta)$, from (4.29) and (4.30) we get that

$$(\widetilde{\theta} - \theta) \quad is \quad AN\left([I(\theta)]^{-1}\,\nu(\theta),\ \frac{[I(\theta)]^{-1}\,M(\theta)\,[I(\theta)]^{-1}}{n}\right) \tag{4.31}$$

which is equivalent to (4.28). This completes the proof of the theorem. □

Since $I(\theta)$ is the limit of (4.25), from the above theorem and using (4.24), we can see that the first stage quasi-least squares estimate $\widetilde{\beta}$ is a consistent estimate of β, whereas $\widetilde{\alpha}$ and $\widetilde{\gamma}$ are

asymptotically biased. To get consistent estimates of α and γ we will make a transformation on $\tilde{\theta}$, which depends on the structures of the correlation matrices $R_T(\alpha)$ and $R_P(\gamma)$. Let

$$b(\overline{\theta}, \theta) = (b_1(\overline{\beta}, \beta), b_2(\overline{\alpha}, \alpha), b_3(\overline{\gamma}, \gamma))' \tag{4.32}$$

be a function of $\overline{\theta} = (\overline{\beta}, \overline{\alpha}, \overline{\gamma})'$ and $\theta = (\beta, \alpha, \gamma)'$, where

$$\begin{aligned} b_1(\overline{\beta}, \beta) &= \overline{\beta} - \beta \\ b_2(\overline{\alpha}, \alpha) &= \mathrm{tr}\left(\left. \frac{\partial R_T^{-1}(\alpha)}{\partial \alpha} \right|_{\alpha=\overline{\alpha}} R_T(\alpha) \right) \\ b_3(\overline{\gamma}, \gamma) &= \mathrm{tr}\left(\left. \frac{\partial R_P^{-1}(\gamma)}{\partial \gamma} \right|_{\gamma=\overline{\gamma}} R_P(\gamma) \right). \end{aligned} \tag{4.33}$$

Note that the second stage quasi-least squares estimate is the solution $\widehat{\theta} = \kappa(\tilde{\theta})$, of the equation $b(\tilde{\theta}, \theta) = 0$. The next theorem shows that $\widehat{\theta}$ is a consistent estimate of θ and asymptotically normal.

THEOREM 4.2 *Let $\theta = (\beta, \alpha, \gamma)'$ be fixed. Let $\tilde{\theta}$ be as in Theorem* 4.1 and $b(\tilde{\theta}, \theta)$ be as defined in (4.32). *Assume that the conditions of Theorem* 4.1 hold. *Let $\widehat{\theta} = (\widehat{\beta}, \widehat{\alpha}, \widehat{\gamma})'$ be the solution, say $\kappa(\tilde{\theta})$, of the equation $b(\tilde{\theta}, \theta) = 0$, where $\kappa(\cdot)$ is a continuous function. Then $\widehat{\theta}$ is a consistent estimate of θ and $\sqrt{n}\,(\widehat{\theta} - \theta)$ converges in distribution to a normal distribution with mean 0 and covariance matrix*

$$\Gamma(\theta) = \left[\nabla\kappa(\theta^*)' \, [I(\theta)]^{-1} \, M(\theta) \, [I(\theta)]^{-1} \, \nabla\kappa(\theta^*) \right] \tag{4.34}$$

where $\theta^* = \theta + [I(\theta)]^{-1} \, \nu(\theta)$ and $\nabla\kappa(\theta^*)$ is $\partial\,\kappa(\theta)/\partial\,\theta'$ evaluated at $\theta = \theta^*$. Finally, $\widehat{\phi}$ as defined in (2.11), is a consistent estimate of ϕ.

Proof: From Theorem 4.1, we know that $\tilde{\theta} = (\tilde{\beta}, \tilde{\alpha}, \tilde{\gamma})'$ converges to $\theta^* = (\beta^*, \alpha^*, \gamma^*)'$ as $n \to \infty$. Note that $\beta^* = \beta$. Using the weak law of large numbers, we can check that

$$U(\tilde{\beta}, \tilde{\gamma}) = \frac{1}{n\,p} \sum_{i=1}^{n} Z_i'(\tilde{\beta}) \, R_P^{-1}(\tilde{\gamma}) \, Z_i(\tilde{\beta}) \;\to \frac{\phi}{p} \, \mathrm{tr}(R_P^{-1}(\gamma^*) \, R_P(\gamma)) \, R_T(\alpha) \tag{4.35}$$

and

$$V(\tilde{\beta}, \tilde{\alpha}) = \frac{1}{n\,t} \sum_{i=1}^{n} Z_i(\tilde{\beta}) \, R_T^{-1}(\tilde{\alpha}) \, Z_i'(\tilde{\beta}) \;\to \frac{\phi}{t} \, \mathrm{tr}(R_T^{-1}(\alpha^*) \, R_T(\alpha)) \, R_P(\gamma). \tag{4.36}$$

Taking the limit as $n \to \infty$, using (4.35) and (4.36) we get

$$0 = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} h_i(\tilde{\theta}) = \begin{bmatrix} 0 \\ \frac{\phi}{p} \operatorname{tr}(R_P^{-1}(\gamma^*)\, R_P(\gamma))\, b_2(\alpha^*, \alpha) \\ \frac{\phi}{t} \operatorname{tr}(R_T^{-1}(\alpha^*)\, R_T(\alpha))\, b_3(\gamma^*, \gamma) \end{bmatrix} \tag{4.37}$$

where the functions b_2 and b_3 are defined in (4.33). Since $\beta^* = \beta$, from (4.37) we can see that $b(\theta^*, \theta) = 0$. Thus $\theta = \kappa(\theta^*)$, and therefore $\widehat{\theta} = \kappa(\tilde{\theta})$ converges to θ. This establishes consistency of the second stage quasi-least squares estimate $\widehat{\theta}$. By the delta theorem it follows that $\sqrt{n}\,(\widehat{\theta} - \theta)$ converges in distribution to a normal distribution with mean 0 and covariance matrix $\Gamma(\theta)$. Finally consistency of $\widehat{\theta}$ implies that $\widehat{\phi}$ is a consistent estimate of ϕ. This completes the proof of the theorem. □

REMARK 4.1 Since $M_{n11}(\theta) = \phi\, I_{n11}(\theta)$, taking limit as $n \to \infty$, using (4.27) and Theorem 4.2, we can see that $\sqrt{n}\,(\widehat{\beta} - \beta)$ converges to a q-variate normal distribution with mean 0 and covariance matrix $\phi\,[C(\theta)]^{-1}$, where

$$C(\theta) = \lim_{n\to\infty} \left[\frac{1}{n} \sum_{i=1}^{n} D_i'(\beta)\, \Sigma_i^{-1}(\theta)\, D_i(\beta) \right]. \tag{4.38}$$

Thus $\widehat{\beta}$ is asymptotically an efficient estimator. The same asymptotic property also holds for the estimate of β, obtained solving the equation (2.6) after substituting $\widehat{\alpha}$, $\widehat{\gamma}$ for α and γ respectively.

4.2 Test of Hypothesis

Suppose that we are interested in testing the null hypothesis $K'\beta = m$, where m is a known $s \times 1$ vector and K is a known $q \times s$ matrix of rank $s \le q$. We propose the test statistic

$$T_n = \frac{(K'\widehat{\beta} - m)'\,(K'(\sum_{i=1}^{n} D_i'(\widehat{\beta})\, \widehat{\Sigma}_i^{-1}\, D_i(\widehat{\beta}))^{-1}\, K)^{-1}\,(K'\widehat{\beta} - m)}{\widehat{\phi}}. \tag{4.39}$$

where $\widehat{\Sigma}_i = \Sigma_i(\widehat{\theta})$ and $\widehat{\phi}$ is defined in (2.11). Large values of T_n are considered to be significant. It can be seen easily from Theorem 4.2 and Remark 4.1 that, under the above null hypothesis, T_n converges to a central χ^2 with s degrees of freedom. We will use the test statistic T_n to test various hypothesis in the examples considered in Section 5.

5 Examples

To illustrate the estimation of the regression and various correlation and scale parameters (by implementing the algorithm described in Section 2.3) and to perform certain hypotheses testing, in this section we present the analyses of two real life data sets. Both the examples have three response variables measured over three time periods. For both the examples we have fit general correlation structure for the three response variables and an AR(1) structure for the measurements observed over the three time periods. In the first example there is only one group, whereas in the second there are two groups.

5.1 AIDS Data

Here we consider the data set given in Table 1 of Thompson (1991). Twenty seven patients were involved in a pilot study where a new drug was being tested for treating AIDS. Measurements on three variables ($p = 3$): TMHR score, Karofsky score, and T-4 cell count, were observed on each of the 27 ($n = 27$) patients at three time periods ($t = 3$), in the beginning, 90 days after the treatment, and 180 days after the treatment. We fit a regression model for these data with the correlation structure $[R_T(\alpha) \otimes R_P(\gamma)]$, where $R_T(\alpha)$ is the matrix of AR(1) correlation structure and $R_P(\gamma) = R_P$ is the unstructured correlation matrix. In order to achieve this that the variance of each variable is approximately equal, we divide each response variable by its sample standard deviation (actually a value close to it). For the present example, we divided the observations corresponding to each of the three variables respectively by 2.4, 12.6, and 276.0. Interest is to test the effect of the drug over time. Hence the null hypothesis we want to test is that there is no effect over time for each of the three variables. As a preparation for testing this hypothesis, suppose y_{ijk} is the observation on the jth variable taken at the kth time period corresponding to the ith individual. Then consider the model

$$E(y_{ijk}) = \mu_{jk},\ j = 1,2,3;\ k = 1,2,3;\ \text{and } i = 1,...,27.$$

or $E(y_i) = X_i\beta = \mu$, where $y_i = (y_{i11}, y_{i21}, y_{i31}, ..., y_{i33})'$ and $\mu = (\mu_{11}, \mu_{21}, ..., \mu_{32}, \mu_{33})'$. Then the parameter estimates obtained using our algorithm are:

$$\widehat{\mu} = (2.1836, 6.3198, 1.1770, 0.8488, 7.3486, 1.2093, 0.9259, 7.6426, 1.0645)',$$

$$\widehat{R}_P = \begin{bmatrix} 1.0000 & -0.5124 & -0.4687 \\ -0.5124 & 1.0000 & 0.4700 \\ -0.4687 & 0.4700 & 1.0000 \end{bmatrix},\ \widehat{\alpha} = 0.6696,\ \text{and } \widehat{\phi} = 0.8760.$$

The null hypothesis of interest then can be expressed as

$$H_0:\ \mu_{j1} = \mu_{j2} = \mu_{j3},\ \text{for all } j = 1,2,3.$$

or $H_0 : K'\mu = 0$, with

$$K' = \begin{bmatrix} 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \end{bmatrix}_{6\times 9}$$

To test H_0 we use the test statistic (4.39) and the observed value of this is 161.1373. The P-value using the chi-square distribution with 6 degrees of freedom is 0.0000. Thus rejecting the null hypothesis.

Next we want to see whether the change in the patient's condition occurred during the first 90 days and/or the second 90 days. For that we test the two hypotheses $H_0 : K_1'\mu = 0$ and $H_0 : K_2'\mu = 0$, where

$$K_1' = \begin{bmatrix} 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 \end{bmatrix}$$

and

$$K_2' = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \end{bmatrix}.$$

The test statistic for the first 90 days is 140.3697 with a P-value using the chi-square approximation on three degrees of freedom is 0.0000. The test statistic for the second 90 days is 11.7403 with a P-value of 0.0083. Thus our analysis shows a significant change in both the time periods. It is easy to check by performing these analyses for each variable separately that there is no significant change in T-4 cell counts in any one of the two time periods, there is a significant change in TMHR score in the first 90 days time period only, and there is a significant change in Karofsky score in both the time periods.

One last thing we want to determine in this analysis is whether the change detected is in the right direction. For an improving patient the TMHR score should decrease, the Karofsky score should increase and the T-4 cell count should increase as well. For determining this we fit the following linear model for the mean:

$$\mu_{jk} = \beta_{0j} + \beta_{1j}x, \ x = 1, 2, 3, \text{ and } j = 1, 2, 3.$$

For an improving patient we want $\beta_{11} < 0$, $\beta_{12} > 0$ and $\beta_{13} > 0$. Fitting this model yields $\widehat{\beta}_{11} = -0.6289 < 0$, $\widehat{\beta}_{12} = 0.6614 > 0$, which have the correct signs for indicating an improvement. However, $\widehat{\beta}_{13} = -0.0562 < 0$, indicating that the improvement is not in the correct direction.

But as mentioned above, an analysis of T-4 cell counts had shown that there is no significant change in this cell counts. Since this was an experimental drug, it is possible that it was not effective in controlling the AIDS virus, from all perspective.

5.2 Zullo's Dental Data

To further illustrate testing of various hypotheses, we use Zullo's dental data appeared in Table 7.2 of Timm (1980). These data were also analyzed by Naik and Rao (1997) assuming a Kronecker product structured covariance matrix for the covariance between the observations on an individual, but using maximum likelihood theory.

The study was concerned with the relative effectiveness of two orthopedic adjustments of the mandible. Nine subjects were assigned to each of the two orthopedic treatments, say T_1 and T_2 ($g = 2, n_1 = 9, n_2 = 9$), called activator treatments. The measurements were made on three characteristics ($p = 3$), namely, SOr-Me (in mm), ANS-Me (in mm), and Pal-MP angle (in degrees) to assess the changes in the vertical position of the mandible at three time points ($t = 3$) of activator treatment. The three null hypotheses of interest are: there is no group and time interaction, there is no group effect and there is no time effect.

Suppose y_{ijkl} is the observation on the kth variable at the lth occasion corresponding to the ith individual in the jth group. We assume the following model for the expected value $E(y_{ijkl}) = \mu_{jkl}$:

$$\mu_{jkl} = var_k + group_{jk} + time_{kl} + (group * time)_{jkl}.$$

To express the above model in the standard form as $E(y_i) = X_i\beta$, we first divide the observations corresponding to the three variables, SOr-Me, ANS-Me, and Pal-MP angle by 7.34, 4.76, and 5.56 respectively. Next we define the following dummy variables:

$$x_{v1} = \begin{cases} 1 \text{ if the observation is on variable 1} \\ 0 \text{ otherwise,} \end{cases}$$

$$x_{v2} = \begin{cases} 1 \text{ if the observation is on variable 2} \\ 0 \text{ otherwise, and} \end{cases}$$

$$x_{v3} = \begin{cases} 1 \text{ if the observation is on variable 3} \\ 0 \text{ otherwise.} \end{cases}$$

The coefficient of each of these in the model will represent the unstructured mean of that variable. Next let

$$x_g = \begin{cases} 1 \text{ if the observation is from group } T_2 \text{ and} \\ -1 \text{ if it is from } T_1. \end{cases}$$

Testing that the coefficient of x_g in the model is zero will test the hypothesis that there is no group effect. To test for the time effect, let

$$x_{t1} = \begin{cases} 1 \text{ if the time period is two} \\ -1 \text{ if the time period is one} \\ 0 \text{ otherwise, and} \end{cases}$$

$$x_{t2} = \begin{cases} 1 \text{ if the time period is three} \\ -1 \text{ if the time period is one} \\ 0 \text{ otherwise.} \end{cases}$$

Now the null hypothesis of no time effect is same as testing that the regression coefficients corresponding to x_{t1} and x_{t2} are simultaneously zero. Finally, the no interaction hypothesis can be tested by testing that the regression coefficients corresponding to the products $x_g * x_{t1}$ and $x_g * x_{t2}$ are zero. The estimates of the parameters corresponding to these eight independent variables are

$$\widehat{\beta} = (16.7591, 13.6798, 4.4253, 0.0385, 0.1674, 0.1105, -0.0032, 0.0009)'.$$

With the assumption $Cov(y_i) = \phi(R_T(\alpha) \otimes R_P(\gamma))$, where $R_T(\alpha)$ is the matrix of AR(1) correlation structure and $R_P(\gamma) = R_P$ is the unstructured correlation matrix, the estimates of the correlation parameters are:

$$\widehat{R}_P = \begin{bmatrix} 1.0000 & 0.7478 & 0.0264 \\ 0.7478 & 1.0000 & 0.3364 \\ 0.0264 & 0.3364 & 1.0000 \end{bmatrix}, \quad \widehat{\alpha} = 0.9381, \text{ and } \widehat{\phi} = 0.9678.$$

The value of the test statistic for testing no interaction is 0.0136, indicating no interaction between the treatment groups and the time period. Similarly test for testing no group effect also showed no significance with a value of test statistic to be 0.4861. Only time effect is significant with test statistic value 23.8208 and the corresponding P-value based on chi-square distribution with two degrees of freedom is 0.0001.

6 Concluding Remarks

In this paper we discussed the analysis of multivariate repeated measures data assuming that the covariance matrix of the repeated measurements on each subject is a scale multiple of Kronecker product of two correlation matrices. The method used is the quasi-least squares, which does not make any assumptions on the distribution of the random errors except for the existence of the first two moments. We have suggested an algorithm for computing the estimates for finite samples. And proved consistency and asymptotic normality of the estimators for large samples and suggested tests for testing any linear hypothesis. Finally we have implemented these results on two real life data sets. Since the quasi-least squares method uses the solution for a set of best (optimal if the data are normal) unbiased estimating equations, it is one of the best procedures for analyzing these data without making any distributional assumptions.

7 References

1. Boik, J. B. (1991). Scheffe''s mixed model for multivariate repeated measures: A relative efficiency evaluation. *Commun. Statist.-Theory Meth.* **20**, 1233-1255.

2. Chaganty, N. R. (1997). An alternative approach to the analysis of longitudinal data via generalized estimating equations. *J. Statist. Plann. Inference* **63**, 39-54.

3. Chaganty, N. R. and J. Shults (1999). On eliminating the asymptotic bias in the quasi-least squares estimate of the correlation parameter. *J. Statist. Plann. Inference* **76**, 145-161.

4. Crowder, M. (1995). On the use of a working correlation matrix in using generalized linear models for repeated measures. *Biometrika*, **82**, 407-410.

5. Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Stat.*, **31**, 1208-12.

6. Heyde, C. C. (1997). *Quasi-Likelihood And Its Applications.* Springer-Verlag, New York.

7. Khattree, R. and D. N. Naik (1999). *Applied Multivariate Statistics with SAS Software,* 2nd ed. SAS Institute, Cary and Wiley, New York.

8. Liang, K-Y. and S. L. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13-22.

9. Naik, D. N. and S. Rao (1997). Analysis of multivariate repeated measures data with a Kronecker product structured covariance matrix. *Preprint,* Dept. of Math. and Stat., Old Dominion University, Norfolk, VA.

10. Rao, C. R. and M. B. Rao (1998). *Matrix Algebra and Its Applications to Statistics and Econometrics*, World Scientific, Singapore.

11. Serfling, R. J. (1981). *Approximation Theorems of Mathematical Statistics*, John Wiley and Sons, New York.

12. Shults, J. and N. R. Chaganty (1998). Analysis of serially correlated data using quasi-least squares. *Biometrics* **54**, 1622-1630.

13. Thompson, G. L. (1991). A unified approach to rank tests for multivariate and repeated measures designs. *J. Amer. Statist. Assoc.* **86**, 410-419.

14. Timm, N. H. (1980). Multivariate analysis of variance of repeated measurements, *Handbook of Statistics*, Vol.1, ed. P. R. Krishnaiah, pp. 41-87, North-Holland, New York.